Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2022-70-RC2, 2022

Comment on gmd-2022-70

Anonymous Referee #2

Referee comment on "The AirGAM 2022r1 air quality trend and prediction model" by Sam-Erik Walker et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-70-RC2, 2022

General comments

The authors proposed a statistical model – AirGAM – to estimate trends of daily concentrations of several air pollutants observed from air quality monitoring stations (AirBase/AQER data) when the effects attributable to meteorology and time are properly accounted for. This process is called or similar to "meteorological normalization" in other publications. As its name suggests, the model is based on the generalized additive model (GAM) technique, a non-parametric regression model capable of fitting the nonlinear relationship between concentrations and predictors (meteorological and time covariates) considered in this study. The airGAM model incorporated several well-established R-libraries (mgcv, openair, and sandwich) for performing most of statistical analyses and visualization.

For reader who are familiar with GAM or have used GAM, the idea of the airGAM would be intuitive. GAM is used to fit time series of the daily concentration of an air pollutant, based on both meteorological and time covariates. A trend disaccounting for the effects of trends of meteorological variables and time variations (weekday and day of the year) represents the meteorology-adjusted trend. In general, it is a smart idea of performing the trend analysis using the approach presented in this work, which I was not aware of or did not come up with (this statement may be biased).

However, the manuscript is interpreted poorly. The authors put too much effort in explaining the trivial model configurations and features (Section 3-5), which distracts readers from focusing the essential contribution and results of the airGAM model, disrupting the readability of the manuscript.

Based on the current status of the code, there is a huge potential to improve the general usability and user-friendliness of the model. For users who are not very proficient in R,

they might be scared away by the tedious operations on data curation and model configuration. For experienced R users, once they got the idea, they could perform/customize the same analysis using mgcv library, rather than sticking to the airGAM script. I would encourage the authors to write at least a user-friendly manual of the model.

Line specific comments:

[L. 22] 'thought to be'? Does it mean that the emissions and background concentrations (e.g., inventory data?) can be derived implicitly from the non-linear relationship between meteorology, time, and concentrations? I am curious to which extent this statement is true.

The emissions data are often provided at monthly/annual intervals. This could be another reason for not including the emissions data into the model.

[63] Typo. The citation might be 'Chang et al. (2021)'

[L.116] "these are thought"?

[L. 211, Table 1] Does UT mean "universal time"? As the AirBase/AQER stations are scattered across several time zones, I would doubt whether it is appropriate to use variables at 18 UT. The atmospheric stability may differ substantially for stations located in different time zones.

From the context, it is not clear which ERA5 reanalysis data were used in the model. But based on the references, it is the "ERA5 hourly data on single levels from 1979 to present" that were used.

As far as I know, ERA5 hourly data on single levels do not provide humidity variables directly. If the humidity variables are derived using 2m air temperature, 2m dewpoint temperature, and surface pressure, please describe the formulae used.

10m wind direction is not directly given in the dataset neither (at least not for land areas).

[Table 1] Are the cyclic variables (wind direction, dayofyear) transformed in the model? For example, day 1 and day 365 diverge remarkably in value but are very close in time.

[L. 271 - 272] it is not clear for me what the sentence "Since ….A= Y_bar" means.

[Section 2.2.2] If the select can be set as "TRUE", incorporating background emissions into the model may not be a problem. If it does not bring benefits to add emission data, GAM can detect and delect it.

[L. 399] Please justify the usage of sandwich library and vcovHAC routine.

[Section 3] For better readability I would suggest that the description on technical details be put into the supplement (in the form of a manual). If the authors would move a step further to demonstrate the usage of the AirGAM model and introduce various functions offered by the package, a tutorial made by Rmarkdown would be helpful.

[L.543] According to the definition of winter and summer made here, the entire year is divided into two parts – winter (oct-mar) and summer (apr-sep). This contradicts Lines 567-568 where each season comprises three months. Consider changing to warm-cool, hot-cold?

[L. 611] "The default setting of this variable, when the trend_type is nonlinear, is NA, …"

[Section 4-5] These two sections could be also put into the supplement.

[L. 961] "blue curve". Please try to give quantitative measures (e.g., MSE, R2, Nash-Sutcliffe Efficiency) to support the statement of good correspondence between modelled and observe values.

[Section 5.3] The section on the probabilistic model evaluation needs to be expanded to include more details on the performance measures adopted. It is not intuitive how these metrics are defined, even the references are given.

[L. 1384] Can you add an inlet figure (e.g., percentage change vs. longitude) to support

the west-east decreasing gradient?

[L. 1464-1465] It’s not clear why the trend analysis was not conducted for stations in France. Are the French data provided at hourly or daily interval? If the French data are hourly, they could be easily aggregated to daily to perform the trend analysis.

[L. 1552] Why not using the same dataset with the same temporal coverage to compare results from the two approaches?

[L.1859] typo, ‘PM2.5’

[Appendix A] I do not think the details on how to install R would be necessary. As far as I can see from the Zenodo repository, the AirGAM model has not been wrapped into an independent library, but it is run like a normal script in R/Rstudio. For better user-friendliness, it would be nice if the model is provided as a R-library with usage examples.