EGUsphere, referee comment RC1
https://doi.org/10.5194/egusphere-2022-396-RC1, 2022

Comment on egusphere-2022-396

William Stockwell (Referee)

Referee comment on "Predicting peak daily maximum 8□h ozone and linkages to emissions and meteorology in Southern California using machine learning methods (SoCAB-8HR V1.0)" by Ziqi Gao et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-396-RC1, 2022

General Comments

This paper presents four different empirical models for estimating peak daily maximum 8-hour (averaged) ozone from meteorological factors and the level of nitrogen oxide emissions (NOx = NO2 + NO). Four different statistical models: the Generalized Additive Model (GAM), the Multivariate Adaptive Regression Splines, the Random Forest, and the Support Vector Regression were developed and applied to estimate ozone concentrations in the South Coast 25 Air Basin (SoCAB) of California, including Los Angeles and the surrounding region.

The use of empirical models for estimating extreme ozone concentrations is particularly relevant because these results may be of interest for health and regulatory purposes. The models may be improved further as the available datasets become larger due to more observations being made over time. Empirical / statistical models are usually more accurate than first-principles numerical forecast models (as long as there are no large changes in the conditions used to derive the empirical models).

There is a long history of the development of empirical models that extends back several decades. However, there are new concepts in machine learning that the authors have used to inform their research. I commend the authors for their appropriate citation of the literature, but they might consider a paragraph to mention the long history of empirical / statistical models.

Specific Comments

The authors provide an excellent discussion of their four models: the Generalized Additive Model (GAM), the Multivariate Adaptive Regression Splines, the Random Forest, and the Support Vector Regression. This clearly written presentation is an outstanding introduction to modern empirical modeling. I can easily imagine using this paper in a graduate atmospheric science course.

The correlations between the model predictions and observations are high and the biases are low. There are only small differences in the performance, in terms of accuracy and required computational resources, between the four approaches for the dataset examined. It would be interesting to see a similar comparison for a much larger dataset in a future paper. Overall, I find that this paper by Gao et al. to be a valuable contribution to the literature.

Technical Corrections

Please consider a paragraph to mention the long history of empirical / statistical models if space allows.