Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2022-134-RC1, 2022

Comment on gmd-2022-134

Anonymous Referee #1

Referee comment on "Development of a regional feature selection-based machine learning system (RFSML v1.0) for air pollution forecasting over China" by Li Fang et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-134-RC1, 2022

The paper is quite interesting, but I still have some comments on it

The major comments: 1. how could you identify the improvement comes from your new methods or just spliting data to 6 groups? spliting data into spatial groups also can help model more easier to capture the variation.

2.why you split the sites into 6 categories? and when you apllied your final models, how could you define the predicted location/grid belong to which categories?

3.please add the spatial cross-validation results to check your model spatail predict ability

4.for the temporal validation, .the test data are only in winter? have you tried to use rolling temopral validation? use previous 4 seasons as training, 1 season as validation.

5.The study only use 2 year data, I wonder whether model can be predicted in the following year.

6.the study period did not include 2020, so will the model forecast be affected by covid-19 when we applied this model for the real early warning system? Have you test the models with 2020 or 2021? this will be the important issue for early warning system.

7.the ground-level pollutant is also included as input? how could you predict the value for thoes location outside these sites?

Minior comments

1.why the max depth of RF is Unknown?

2.why each region choose 15 sites for features selection? have you tried 10,20, or some other numbers?

3.why do you pick only top 3 features? not 5 or more? will they affect the results?

Technique points:

1."Cubic imputation was applied to fill in the missing data because the 3 h resolution of the CAMS reanalysis data is too coarse for interpolation. After interpolation, 0.75o x 0.75o grid data were imputed to each monitoring station" here you mean impute the missing or downscale? if downscale, from 0.75 to which resolution?

2."The data were interpolated to the monitoring station locations for use in machine learning." It is still confusing.