Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-200-AC1, 2021

Geoscientific Model Development Discussions

Interactive comment on "Machine learning models to replicate large-eddy simulations of air pollutant concentrations along boulevard-type streets" *by* Moritz Lange et al.

Moritz Lange et al.

kai.puolamaki@helsinki.fi

Received and published: 4 June 2021

We thank the referees for their encouraging comments.

Regarding the specific comments by the referees:

RC1, scaling variable: The reason for the scaling variable is, as stated in the manuscript (page 9), that the scaling - including units of measurements - of air pollutant concentrations in the evaluation data (KA20) differ from those in the training data (KU18). For this reason, we use an error measure (mRMSE) that only depends on the relative "shape" of the pollutant concentration and that is invariant to linear scaling

of the concentrations in the training and/or evaluation data. Because the scaling of the training and evaluation data differ it is not possible to compute this scaling factor by using the training data alone, which is why we compute the scaling variable on the evaluation data (KA20) by finding the scaling which minimises the RMSE on the evaluation data.

For new evaluation data set we could either use the same multiplicative constant (if the scaling in the new evaluation data set is expected to be identical the scaling in the old evaluation dataset) or find a new multiplicative constant.

(An alternative interpretation would be that the scaling variable would be part of our regression model and that we would use the standard RMSE error measure, in which case we would make a (small) error when we find the scaling constant by using the evaluation data. Even with this alternative interpretation, the resulting overfitting should however not be substantial, because we are fitting here only one number to the evaluation data.)

We will add a short discussion about these issues to the description of mRMSE error measure.

RC2, comment 1: The reviewer is correct in pointing out that we could solve the problem of negative outputs, e.g., in the simple linear models by transforming the output variable to positive values. In fact, this is what is effectively accomplished by the Poisson and log-linear regression models which transform the linear response, e.g., by using exponential transformation to non-negative domain and which do not therefore have the problem of negative outputs. We have included the linear model in the manuscript, because we just wanted to include "naive" linear regression as a simple baseline model. We will clarify this in the final manuscript.

RC2, comment 2: As pointed out in page 11 of the manuscript, standard SVR leads to practically similar performance with logarithmic SVR. Out of these two SVR models we choose to highlight the latter (logarithmic SVR) because it has always non-negative

predictions and because the performance and chosen features are otherwise practically identical with the standard SVR.

RC2, comment 3: Reporting the RMSE for the dummy model allows to assess the overall benefit of using a regression model at all. For example, knowing (m)RMSE for the dummy model allows to easily estimate the commonly used Rˆ2 metric (e.g., Rˆ2=1-0.76ˆ2/1.78ˆ2=0.82 for the log-linear regression mentioned in the abstract). For these reasons, we feel that reporting the dummy model RMSE as a baseline is informative.

Sincerely,

The Authors

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-200, 2020.