Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-243-RC2, 2021

Earth System Science Data Discussions

Interactive comment on "Long-term trends of ambient nitrate (NO_3^-) concentrations across China based on ensemble machine-learning models" *by* Rui Li et al.

Anonymous Referee #2

Received and published: 18 February 2021

In this study, Rui et al., attempt to use an ensemble of models to provide spatially and temporally resolved particulate nitrate concentrations in China. Of course, the availability of a reliable nitrate estimation across China would have immense value, but I see a number of issues in the presentation of the material. The sub-models are all poorly described and frankly quite confusing. The authors should clarify that they understand these beyond simply clicking the necessary buttons in a software package. Moreover, use of the explanatory variables is muddled. We don't know which variables were used and which thrown out due to redundancy, or how consistently that was done across sub-models or time periods. Finally the MLR statistics that define the ensemble model are not provided, including no discussion of multicollinearity issues. Application

of the ensemble model that followed is of course only relevant if the model itself is well established.

Based on Figure S1, the site in far northwest China only seems to have data from 2010-2011. This may dramatically impact the reliability of the method as that site is singularly located and likely unique in terms of the meteorological variable range, which the models are dependent on. Please comment on this data issue

Please provide information in the SI material and comment in the text about the monthly sample size of useable NO2 column data.

Sec 2.1: (i) These are described as monthly samples. Please clarify that they are continuously collected over a month and not a single 24-h sample each month. I believe the former is true, correct? (ii) you should include the reference to Xu et al., (2019) which reports NNDMN data from 2010-2015, (iii) within the Xu et al., (2019), they specify that the NO3 data is in units of ugN/mˆ3 and I wonder if you corrected for the molecular weights in this regard when you reported as NO3 concentrations? Your values seem similar to the ugN/mˆ3 values from Xu et al., (2019); (iv) finally, please specify the PM size-cutoff. This will impact whether nitrate from soil dust influenced the measurements

Xu, W., Zhang, L. & Liu, X. A database of atmospheric nitrogen concentration and deposition from the nationwide monitoring network in China. Sci Data 6, 51 (2019). https://doi.org/10.1038/s41597-019-0061-2

Sec. 2.3: with all of the variables introduced here, were they all incorporated into the MLR or just the sub-models? In the last paragraph of the section you mention, they are all incorporated into the model, but you don't indicate which model(s). Please clarify. And please provide detailed information about the finalized variables and which were found to be redundant. And was this made to be consistent for each time period? or the final predictor variables varied with time period analyzed?

Sec. 3.1 Equations 1-5: I (capital i), L, and chi are all not defined and generally make the use of this model too murky to understand

Equation 6: R is not defined; c would be better to have a subscript, perhaps naught?; what is "arg"?

Equation 8: again, undefined symbols and a general lack of description do not give the reader much confidence that the authors understand the complexities of the models that they are using

Sec. 4.2: I don't see the MLR statistics anywhere. Coefficients? Standard error of the coefficients? The MLR should also indicate whether or not it was really helpful to stack the models in the first place. You would expect this to suffer from multi-collinearity problems in that all three sub-models are attempting to predict the same nitrate concentrations. The authors can provide a variance inflation factor with interpretation in the text.

Figure 2: shouldn't each of the sub-models be connected with the original data tree (NO2 column / met data / land use) to show their optimization before incorporation into the MLR?

Figure 3: (i) something is missing from the caption. You seem to have forgotten to label what is 3a or 3d, (ii) please include sample size for each plot; (iii) why is there a different sample size for each plot?

Figure 4: (i) I don't understand why there is a color bar...is it for the number of points reported from another study? how can it be between 1 and 2?, (ii) the boxes are different sizes. Does it mean that these represent distributions of predicted NO3 concentrations? How to interpret the regression in this manner?

Figure 5: In the caption, Please specify that these values are from your ensemble model output

Figure 6: again, please specify these are from ensemble model output. "Satellite-

derived" could be mistaken for reanalysis data

Figure 7: (i) are these averaging all grid cells within a certain boundary? Are these boundaries specified anywhere? Please provide details in the text (ii) add the average measured trends for plots b,c, and d

Figure 8: plots c and f are confusing. For most of the map, you show a change of only -0.1 to 0 ug/m3 and with a significant trend!. Are these comparing two annual values each with n=12 at each grid cell? or 24 points in sequence? Even if the latter, it seems incredible that less than a 0.1 ug/m3 change tested significant over 24 points

Table 1: (i) should the first column be labeled "year"? or you intended to add more data to the table? (ii) somewhere you should indicate the sample size for the various years

Figure S1: plots should be labeled with year of data or with a letter and specified in the caption

Figure S3: all abbreviations should be described in the figure caption

Other comments:

Lines 76-79: this sentence is difficult to understand. Please revise

Lines 106-107: please expand this sentence to clarify your meaning or perhaps delete it

Lines 117-119: this sentence is an example where you should specify "particulate nitrate concentration" and there are numerous other places that you should indicate this to distinguish from a possible misinterpretation of rainwater or cloudwater nitrate, the former which is also measured at NNDMN sites

Line 127: NNDMN is not defined

Line 135: this reference should be specified 2018a or 2018b or both

Lines 151-154: this is not a complete sentence, please clarify your meaning

Line 206: "to determine the appropriate parameter" ..... does it mean a regression coefficient? or this is the 'c' value from the various models?

Lines 231-233: what does the first part of this sentence mean? "we only estimated the missing ratio of NO2 column..." please clarify

Lines 242-243: specify these as measured monthly NO3 concentrations

Line 249: I don't see urban land area designations in Fig S2

Lines 252-253: is there anything else to comment about this? any interpretation?

Lines 259-260: I don't understand why you say 'opposite trends' ....doesn't seem opposite. XGBoost is worst and ensemble best in all statistics

Line 269: "Furthermore," should be "However,"

Lines 277-279: I don't understand the latter half of this sentence. Please revise the text

Lines 280-281: is this for all years data combined? please specify

Lines 292-294: this sentence should start with "Although"

Lines 303-304: I don't understand this sentence. Please revise the text

Lines 304-306 and 307-308: Does this explain why RMSE and MAE are poor in these areas? I don't see the definite connection

Lines 306-307: are these areas in Northwest China? please clarify

Lines 330-331: "especially for hindcast of air pollutants" .... you mean compared to the forecast of air pollutants? Seems strange that a forecast model would have a better statistical result

Lines 365-368: the use of "speed" in these sentences is inappropriate. Example sentence revision for one of them: "For instance, the ambient NO3 level in BTH increased

remarkably by 0.13 μg/m3/year from 2005-2014."

Lines 391-400: decrease of SO2 emissions but not NOx emissions can further lead to NO3 increases because of decreased aerosol acidity, which is dictated by SO4 in particulate matter

Line 461: does this mean average available NO2 columns data was only $\sim$43% each month? Or this was a minimum value and only at one site?

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-243, 2020.