Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-270-RC1, 2020

Geoscientific Model Development Discussions

Interactive comment on "Evaluating the use of Facebook's Prophet model v0.6 in forecasting concentrations of NO_2 at single sites across the UK and in response to the COVID-19 lockdown in Manchester, England" *by* David Topping et al.

Anonymous Referee #1

Received and published: 21 December 2020

Topping et al. use a time series forecasting toolbox 'Prophet' to predict NO2 pollution at various UK measurement sites. While I note that the manuscript might have significant potential both scientifically and methodologically, I find its current value for the scientific community difficult to judge. In particular, the manuscript lacks clarity in its presentation in a number of critical aspects. For instance, there is insufficient detail in the description of the methodological approach and of useful benchmarks for their predictions. I therefore recommend major revisions subject to which I might be able to provide a more informed opinion on the quality of the work presented here. As I said,

the issues might be mostly presentational but I would need to see this confirmed.

Major comment:

- I find the current description of the methodology insufficient, especially considering the journal's modelling scope and the importance of the 'Prophet' procedure for this work. All that is said is that 'Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily periodicity' and then the authors point towards a reference and some 'interpretive' parameters. This is simply not enough. A schematic would help and there has to be additional insight into how the algorithm works. I am also not entirely sure what is meant by 'rolling forecasts 30 days in advance' (L64). Do I understand correctly, that you train and cross-validate on three years and then predict every single hour for the next 30 days in advance before moving the training interval 15 days forward in time? What are the predictors in this procedure? I assume it must be meteorological variables and NO2 concentrations of the previous hour, and lagged weekly, daily, and yearly values (L56/57)? From the entire manuscript I could not fully comprehend what your predictions are actually based on. My initial impression was that you actually try to predict NO2 30 days in advance at some point, which I assume cannot possibly be the case? Does, e.g., the prediction at day 20 in the prediction period at noon, know of the true NO2 value just one hour earlier? Please clarify these aspects.

Point-by-point comments:

- Abstract l4-l6: could you be more quantitative or at least indicate somehow what "promising performance" means? When I read this for the first time, I was also confused what 'regional' sources are compared to 'non-local' sources and why BOTH should affect performance negatively. What would be a good setting in comparison to these two settings?

- L10: I find the ‘simplified approach of fitting to derived NO2-per-traffic volume’ approach is introduced somewhat surprisingly and without context. Why is this used? Is that part of one of the models, as indicated by the word ‘despite’ at the beginning of the sentence? Please clarify.
- L16: ‘effective and simple’ relative to what?
- L21: clarify the timescale of predictions you are interested in. Minutes? Simultaneous (from other measurements)? Hours? Days? As I said, I am not entirely sure yet if you actually intend to predict just the update in NO2 concentrations from one hour to the next, or over longer time intervals.
- L30: Could you give examples of what those challenges are?
- L43: packageS such as Scikit-Learn and bracket typos. Keras has no official citation?
- L51: define what you mean by ‘local’ – point/site measurements I suppose?
- L61: here you mention for the first time that you aim to predict NO2 concentrations a month in advance (which could be understood in different ways). I think this should be clarified earlier on, even in the abstract.
- General comments on section 2.1: given how central the Prophet model is to your paper, I would want to see a more detailed description of what happens ‘under the hood’. Currently, and given that I don’t know the method, it seems like a black box to me, which is impression you would certainly want to avoid. A graphic/schematic would help, too. Why should the Prophet model be advantageous over for example LSTMs? Explaining the modeling process would be really important to make the paper more interesting and more accessible.
- L78 typo processes

- L85: TNO database not defined yet. Maybe link website if appropriate?
- L.89-91: without context this doesn’t make sense to me. If there are 2-months periods, what happened to Jan/Feb 2020? I assume you treat 2020 separately due to the lock-down? Maybe worth pointing out already at this stage.
- L.103: make clearer how the different time periods align somewhere in the manuscript (certain years are used for training, others for prediction and EMEP modelling, I suppose, others again to test the effects of the lock-down?).
- L.131: 3. Results. By this point in the manuscript, I am still not sure how you predict NO2, but you already start presenting results. What are the variables you use as predictors (and at what time lags)? How do you cross-validate to avoid overfitting (you mention that something is done in section 2.1, but I feel this is insufficient for a modelling paper)? How do you evaluate model skill? Surely the Pearson correlations shown in Figure 1b are performed on test data and not on training data, i.e. the sequential predictions on predicted months? Your current manuscript simply does not describe this in sufficient detail and clarity in my opinion and I find it difficult to judge the skill of your approach as a result.
- Figure 1b: Please compare these correlations to time series where you simply prescribe a seasonal cycle (smoothed over several days) or a constant value that is representative of the true annual mean observed values. This would provide a much clearer impression of how much better your predictions actually are compared to very basic models. Urban sites I would expect to show less relative (in %) seasonal variation so that this might explain your smaller error there. Furthermore, try the R2-score (coefficient of determination) rather than just Pearson correlation. The former should be better suited to compare time series of this kind because it does not just consider variance but also magnitude of the prediction error.

- L153: I am sceptical that the EMEP model is a useful benchmark here. It simply seems to have a high bias, but how about an empirical model that simply has no mean bias (for instance the seasonal cycle model mention above, derived from actual observations)?
- L175: At this point it is still not clear to me which variables were actually included in the predictions of results section 3.1...
- Figure 7: relabel y-axis with logarithmic scale I suggest.
- L185: I think this approach requires a more detailed explanation. How is this done exactly with Prophet?
- L207: it is also not clear to me how the traffic data is included in the model. If it is a forecast, do you take traffic measures from the previous hour? How is this relevant to the current hour?
- L220ff: suddenly there is a lot of methodological detail. You have long lost me by then though.
- L245: possibly, but then it could also just be a failure of the regression model to perform well in an effectively different environment?

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-270, 2020.