Authors' response to referee comments: Representing chemical history in ozone time-series predictions - a model experiment study building on the MLAir (v1.5) deep learning framework

Felix Kleinert[1,2], Lukas H. Leufen[1,2], Aurelia Lupascu[3], Tim Butler[3], and Martin G. Schultz[1]

[1]Forschungszentrum Jülich GmbH, Jülich Supercomputing Centre (JSC) , Jülich, Germany
[2]Institute of Geosciences, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
[3]Institute for Advanced Sustainability Studies, Potsdam, Germany

Correspondence: Felix Kleinert (f.kleinert@fz-juelich.de)

General statement

We appreciate the constructive comments of the two anonymous reviewers. In the following, we use blue to reply to Reviewer comments. We mark additional changes that the Reviewers do not explicitly state in magenta. We will use the same colours to display changes within the revised manuscript.

1 Answer to Anonymous Referee #1

"The paper is well written. The authors applied an advanced deep learning architecture and conducted a detailed and interesting sensitivity study on the predictability of ozone using the U-net+LSTM. Analyses on the forecast lead time, the added value of upstream history, and the concept drift could be useful for future studies. The inclusion of reference models enhanced the validity of the results. I recommend the publication of the paper with minor revisions."

- General comments
  - At a resolution of 12 km, could you comment on the potential impact of sampling errors, when your model is applied to real observations in the future? How many sites do you have in each grid box?
    The number of pseudo-stations per grid box varies from one up to three pseudo-stations. All in all 247 grid boxes contain exactly one, 35 grid boxes contain two and 5 grid boxes contain three pseudo-station, respectively. Indeed, the pseudo-station locations are not evenly distributed across Germany; therefore, our model 'sees' some regions with better measurement coverage (e.g. south-west) more often than others (north-west). Given that we are using a combination of Convolution and LSTM layers, we do not expect a too large sampling effect. Nonetheless, we want to point out that sampling errors occur and are a general concern in nearly all geospatial DL applications. As shown in manuscript Fig. 9, our model has serious problems representing rare mountainous stations in the south.

In our case, it is more likely that sampling errors related to the seasons have more effects on our results than the sampling errors introduced by pseudo-station locations. We added the following sentence to the data description: Thereby, 247 grid boxes contain exactly one, 35 grid boxes contain two, and five grid boxes contain three pseudo-station, respectively.

- Could you discuss the ozone chemical regime over Germany and the impact of VOC on ozone predictability?
  The ozone production regime varies according to the location (e.g. cities are more likely to be VOC-limited than the rural background) and also with a strong temporal variability. This is for example expressed in the analysis of ozone trends. There is a lot of site-to-site variability in the ozone trends in Europe, with most stations showing a decreased trend for health-related ozone metrics (Fleming et al., 2018). Also, anthropogenic NOx and NMVOC emissions have both decreased in Europe by about 50% since 1990 (European Environment Agency, https://www.eea.europa.eu/data-and-maps/indicators/eea-32-non-methane-volatile-1/assessment-4), making it hard to gauge the contribution of either of these precursors alone to ozone trends. NOx is routinely measured, but NMVOCs are not, or at least not with enough detail to be useful for an analysis of the ozone production regime. Regarding the contribution of VOCs to ozone, Lupaşcu et al. (2022) showed that anthropogenic NMVOC do not contribute significantly to ozone events, but it is rather biogenic VOC. Unfortunately, the BVOC are not routinely measured, so we do not know the trends. The recent review of the Gothenburg Protocol showed that the peak values of ozone are decreasing (see "Scientific information for the review of the Gothenburg Protocol" (available from https://unece.org/environment/documents/2022/07/session-documents/scientific-information-review-gothenburg-protocol), in particular paragraphs 17-20 and Figure 3 showing the higher percentiles), so putting this together, we can likely say that reductions in NOx emissions in Europe have contributed to lower peak ozone. A similar argumentation applies to the complexity of ozone forecasts. We extended the discussion with the arguments presented above.

- Specific points:

  - L161: Inconsistent font of “DataHandler”
    Fixed.

  - L278-280: Is it possible that the training/test performances are different because your NN model is slightly overfitted towards regions with more pseudo-stations?
    All three sets (train, val, test) cover the same pseudo-stations. Therefore, the mean rescaled losses ($64.08\,\mathrm{ppb}^2$, $63.34\,\mathrm{ppb}^2$ and $64.62\,\mathrm{ppb}^2$) are not a result of overfitting towards 'over-represented areas'. However, regional overfitting is more likely on an individual pseudo-station level, as shown in manuscript Fig. 9. Still, we argue that the temporal differences within the three sets are the primary cause of differences in the score.

  - Figure 9: Could you comment on the degraded performance over the coastal regions? Do these coastal sites have something to do with the results of skill scores in Figure 10? For example, the north wind driven by the sea breeze circulation brings less useful information.

As mentioned in the answer above, on an individual pseudo-station level, overfitting towards overrepresented regions is more likely. The total number of coastal pseudo-stations is low compared to the other areas. As we use daily aggregated data, inter-day circulation patterns like the land-sea breeze circulation are averaged out. Only 7.26% of all training samples show a northern upstream sector. For the validation and testing sets, 2.42% and 4.97% of the corresponding samples show a northern upstream sector, respectively. This corresponds to the lowest occurrence of all sectors in all three sets. Thus the northern sector is underrepresented in all sets leading to degraded performance. We have extended the introduction to Fig. 9 (L277) as follows:

Figure 9 shows how the losses are distributed geographically. At first glance, we can identify regions with high MSE ($> 120\,\mathrm{ppb}^2$, yellowish colours) in the mountainous south and in the northern coastal area. The triangles denote the losses of each station separated for the training (left part of the symbol), validation (right part of the symbol) and test (lower part of the symbol) set. When focussing on the set's loss differences at individual pseudo-stations, we can identify the following pattern: the test set's loss is mostly lower than the validation and train loss in Germany's western and southwestern parts. The high individual validation loss in the northern part is directly related to the large temperature anomaly, as shown in Fig. 1. Thus, the feature combination is not explicitly present in the training set resulting in larger discrepancies in forecasts and observations.

- Figure 11-14: How would the learned non-linearities compare with the WRF-Chem CTM? Could you comment on that?

We did not run the WRF-Chem model multiple times to conduct sensitivity analyses of the WRF-Chem model. We updated Figures 11 to 14 by removing the light grey lines that represented the sensitivity of NN3s for each sample. Instead, we now show the daily aggregated WRF-Chem data at t_0 as a scatter plot for each lead time. Thus, the updated figures show the ozone concentration as a function of the variable of interest (e.g. NO2). This is not a full WRF-chem sensitivity analysis but shows how the input variable of interest relates to ozone concentrations within the WRF-chem model. Moreover, we added the following paragraph to the discussion:

Fast et al. (2014) showed that the reduction of all anthropogenic emissions by 50% led to a similar O3 bias in California as for the case without emissions reduction, although a variability of $\sim 5\,\mathrm{ppb}$ can be seen at supersites location. Abdi-Oskouei et al. (2020) performed several sensitivity studies to assess the impact of meteorological boundary conditions and land surface model on modelled O3 concentration, and they showed that these changes led to a minor sensitivity of average ozone concentration ($< 2\,\mathrm{ppb}$). Georgiou et al. (2018) analysed the impact of the different chemical schemes on predicted trace gases and aerosol concentrations and they should that ozone produced by CBMZ-MOSAIC and MOZART-MOSAIC have similar biases ($10.9\,\mathrm{ppb}$ and $11.6\,\mathrm{ppb}$), while the use of RADM2-MADE/SORGAM mechanism led to a bias of $4.25\,\mathrm{ppb}$. The difference in ozone concentrations was attributed to the difference in VOC treatment. Gupta and Mohan (2015) have also analysed the sensitivity of ozone concentration to the choice of chemical mechanism and noted that the CBMZ performed better than the RACM mechanism due to the revised rate of inorganic reaction in the CBMZ mechanism. Mar et al. (2016) found that the absolute concentration of ozone predicted by the MOZART-4 chemical mechanism is up to 20 $\mu\mathrm{gm}^{-3}$ greater than RADM2 in summer, explained by the different representations of VOC chemistry and different inorganic rate coefficients.

- L330: I think your NN3s model slightly outperforms OLS at most of the pseudo-sites?

  Yes, we have rephrased the part for clarity:

  NN3s outperforms the ols model at most of the pseudo-stations that we trained on exactly the same data. Thus, we can conclude that the non-linearity provided by the neural network is essential to extracting meaningful upstream information.

- L340: I see measurement uncertainties are not utilized for now. Can you comment on potential utilization of measurement uncertainties in the future (e.g., via Bayesian NN)?

  Indeed, our presented approach does not explicitly take measurement uncertainties into account. To account for this data uncertainty, an Bayesian neural network approach by replacing the last network layer containing four nodes (one for each lead time) with a distribution layer would be beneficial. Consequently, instead of learning four individual point estimates, one would learn the parameters of a multivariate distribution, resulting in an ensemble prediction with multiple realisations represented by multiple draws from the learned distribution. Moreover, one could explicitly account for model uncertainty by representing the model weights (point estimates) with learnable distributions. The weights would be sampled from the learned weight distributions during each forward path through the network.We added the following paragraph to the new discussion subsection, "Alternative Neural Network Approaches": Furthermore, the application of Bayesian network architectures can help to characterise data and model uncertainties in future studies (for gridded model examples, see, e.g. Sengupta et al., 2020; Sun et al., 2022; Ren et al., 2022)

2 Answer to Anonymous Referee #2

This paper presents an interesting work on the use of machine learning to forecast tropospheric ozone. The current version answers most of the previous review remarks. However, for the sake of completeness, I would raise two questions that could be addressed in the text as "discussion" even though some of them could lead to more experiments in future works.

- Indeed, the first remark is related to the forecast model chosen by the authors, that creates time-series forecasts for each individual air-quality station. As discussed in the text, the experiments show that better results are obtained when using data from surrounding stations rather than a single one, which seems logical as this "neighbourhood" strategy helps capturing the advection ozone. For this reason, it would be interesting to develop a bit more the comparison with models that use spatial data, as for example https://doi.org/10.1007/s10994-020-05944-x that uses 2D matrices of the entire area in the form of images and video frame prediction algorithm. I believe that WRF-Chem outputs can easily be represented as 2D matrices

  We thank the reviewer for the suggestion to put our study into a broader context. We have added the following paragraph that focuses on the alternative field predictions:

As an alternative to predicting ozone concentrations at pseudo-station levels, one could forecast ozone fields. For example, Steffenel et al. (2021) used the PredRNN++ (Wang et al., 2018) model to forecast the Total Column Ozone for the southern part of South America and parts of Antarctica. Gong et al. (2022) recently used convolutional long short-term memory models (Shi et al., 2015, ConvLSTM) and a generative model Stochastic Adversarial Video Prediction (Lee et al., 2018, SAVP) model to forecasts the 2m temperature for a lead time of up to 12 hours over Europe. However, it remains unclear how the existing video frame prediction models cope with the high dimensionality of atmospheric data, especially when focusing on multiple target variables. Attention mechanisms like those proposed in Vaswani et al. (2017) might have an enormous advantage in the earth sciences as those networks - also known as transformers- are encoded by learned contend rather than fixed positions. Querying from content allows for long-time dependencies, whereas classical recurrent architectures have difficulties with long dependencies.

- The second question concerning the datasets. In my opinion, only a single year seems too limited to catch the dynamic of the atmosphere. As training, validation and test subsets cover different months of the year, seasonal variations are certainly present and cannot be correctly expressed. A more clever strategy would be to use datasets from several years (even if only in a reduced number of months per year), and validate with the same period in another year.

  We agree that our chosen time period is very short, and we, therefore, designed the test period in such a way that it covers three months in 2010 that are also present in the training set in 2009. For clarification, we have added the following sentences to the concept drift discussion:

  In order to reduce the concept drift's effect, extending each data set to multiple years would be beneficial in upcoming studies. This would allow the network to operate on more robust data distributions and thus minimise the risk of out-of-sample predictions.

As stated before, answering these remarks is not mandatory but would greatly contribute to the paper completeness, even if just in the form of a discussion

3 Additional Changes

- L49: We updated the pre-print reference (Leufen et al., 2021b) to (Leufen et al., 2022) as the article is now published
- We added the following sentence to the Acknowledgements: Open Access publication funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 491111487

References

Abdi-Oskouei, M., Carmichael, G., Christiansen, M., Ferrada, G., Roozitalab, B., Sobhani, N., Wade, K., Czarnetzki, A., Pierce, R., Wagner, T., and Stanier, C.: Sensitivity of Meteorological Skill to Selection of WRF-Chem Physical Parameterizations and Impact on Ozone Prediction During the Lake Michigan Ozone Study (LMOS), Journal of Geophysical Research: Atmospheres, 125, https://doi.org/10.1029/2019JD031971, 2020.

Fast, J. D., Allan, J., Bahreini, R., Craven, J., Emmons, L., Ferrare, R., Hayes, P. L., Hodzic, A., Holloway, J., Hostetler, C., Jimenez, J. L., Jonsson, H., Liu, S., Liu, Y., Metcalf, A., Middlebrook, A., Nowak, J., Pekour, M., Perring, A., Russell, L., Sedlacek, A., Seinfeld, J., Setyan, A., Shilling, J., Shrivastava, M., Springston, S., Song, C., Subramanian, R., Taylor, J. W., Vinoj, V., Yang, Q., Zaveri, R. A., and Zhang, Q.: Modeling regional aerosol and aerosol precursor variability over California and its sensitivity to emissions and long-range transport during the 2010 CalNex and CARES campaigns, Atmospheric Chemistry and Physics, 14, 10 013–10 060, https://doi.org/10.5194/acp-14-10013-2014, 2014.

Fleming, Z. L., Doherty, R. M., von Schneidemesser, E., Malley, C. S., Cooper, O. R., Pinto, J. P., Colette, A., Xu, X., Simpson, D., Schultz, M. G., Lefohn, A. S., Hamad, S., Moolla, R., Solberg, S., and Feng, Z.: Tropospheric Ozone Assessment Report: Present-day ozone distribution and trends relevant to human health, Elementa: Science of the Anthropocene, 6, 12, https://doi.org/10.1525/elementa.273, 2018.

Georgiou, G. K., Christoudias, T., Proestos, Y., Kushta, J., Hadjinicolaou, P., and Lelieveld, J.: Air quality modelling in the summer over the eastern Mediterranean using WRF-Chem: chemistry and aerosol mechanism intercomparison, Atmospheric Chemistry and Physics, 18, 1555–1571, https://doi.org/10.5194/acp-18-1555-2018, 2018.

Gong, B., Langguth, M., Ji, Y., Mozaffari, A., Stadtler, S., Mache, K., and Schultz, M. G.: Temperature forecasting by deep learning methods, Geoscientific Model Development Discussions, 2022, 1–35, https://doi.org/10.5194/gmd-2021-430, 2022.

Gupta, M. and Mohan, M.: Validation of WRF/Chem model and sensitivity of chemical mechanisms to ozone simulation over megacity Delhi, Atmospheric Environment, 122, 220–229, https://doi.org/10.1016/j.atmosenv.2015.09.039, 2015.

Lee, A. X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., and Levine, S.: Stochastic Adversarial Video Prediction, Tech. Rep. arXiv:1804.01523, arXiv, http://arxiv.org/abs/1804.01523, arXiv:1804.01523 [cs] type: article, 2018.

Leufen, L. H., Kleinert, F., and Schultz, M. G.: Exploring decomposition of temporal patterns to facilitate learning of neural networks for ground-level daily maximum 8-hour average ozone prediction, Environmental Data Science, 1, e10, https://doi.org/10.1017/eds.2022.9, 2022.

Lupaşcu, A., Otero, N., Minkos, A., and Butler, T.: Attribution of surface ozone to NO$_x$ and volatile organic compound sources during two different high ozone events, Atmospheric Chemistry and Physics, 22, 11 675–11 699, https://doi.org/10.5194/acp-22-11675-2022, 2022.

Mar, K. A., Ojha, N., Pozzer, A., and Butler, T. M.: Ozone air quality simulations with WRF-Chem (v3.5.1) over Europe: model evaluation and chemical mechanism comparison, Geoscientific Model Development, 9, 3699–3728, https://doi.org/10.5194/gmd-9-3699-2016, 2016.

Ren, X., Mi, Z., Cai, T., Nolte, C. G., and Georgopoulos, P. G.: Flexible Bayesian Ensemble Machine Learning Framework for Predicting Local Ozone Concentrations, Environmental Science & Technology, 56, 3871–3883, https://doi.org/10.1021/acs.est.1c04076, 2022.

Sengupta, U., Amos, M., Hosking, J. S., Rasmussen, C. E., Juniper, M., and Young, P. J.: Ensembling geophysical models with Bayesian Neural Networks, arXiv:2010.03561 [physics, stat], http://arxiv.org/abs/2010.03561, arXiv: 2010.03561, 2020.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15, pp. 802–810, MIT Press, Cambridge, MA, USA, event-place: Montreal, Canada, 2015.

Steffenel, L. A., Anabor, V., Kirsch Pinheiro, D., Guzman, L., Dornelles Bittencourt, G., and Bencherif, H.: Forecasting upper atmospheric scalars advection using deep learning: an $$O_3$$ experiment, Machine Learning, https://doi.org/10.1007/s10994-020-05944-x, 2021.

Sun, H., Shin, Y. M., Xia, M., Ke, S., Wan, M., Yuan, L., Guo, Y., and Archibald, A. T.: Spatial Resolved Surface Ozone with Urban and Rural Differentiation during 1990–2019: A Space–Time Bayesian Neural Network Downscaler, Environmental Science & Technology, 56, 7337–7349, https://doi.org/10.1021/acs.est.1c04797, _eprint: https://doi.org/10.1021/acs.est.1c04797, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I.: Attention is All you Need, in: Advances in Neural Information Processing Systems, edited by Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., vol. 30, Curran Associates, Inc., https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf, 2017.

Wang, Y., Gao, Z., Long, M., Wang, J., and Yu, P. S.: PredRNN++: Towards A Resolution of the Deep-in-Time Dilemma in Spatiotemporal Predictive Learning, in: Proceedings of the 35th International Conference on Machine Learning, edited by Dy, J. and Krause, A., vol. 80 of *Proceedings of Machine Learning Research*, pp. 5123–5132, PMLR, https://proceedings.mlr.press/v80/wang18b.html, 2018.