RC1

Review of Global fields of daily accumulation-mode particle number concentrations using in situ observations, reanalysis data and machine learning by Ovaska et al.

Accumulation mode aerosols are important climatologically because of their interactions with radiation and clouds. Measurements of accumulation mode number concentrations from satellites are inferred from radiative properties (e.g., extinction) and only available in cloud-free regions which reduces our ability to constrain aerosol-cloud interactions in climate models. Observations from ground or airborne instruments are spatially and temporally limited but of high fidelity and useful for testing satellite retrievals and/or model simulations. The rise of complex machine learning (ML) approaches offers the opportunity to create diverse datasets that relate observed aerosol number concentrations to more widely observed/simulated meteorological phenomena. In this paper, Ovaska et al use two established explainable-ML approaches to relate observed accumulation number (N100) concentrations to coincident reanalysis fields and thereby create a high-coverage N100 dataset. The methodology accounts for various confounding factors which may bias results including the paucity of surface measurements which are mostly Europe-confined, and the varying-lengths of observational datasets.

The paper is very well written, timely, important and a provides a benchmark method for calculating aerosol number concentrations from predictor variables using ML methods. As someone with limited ML knowledge I particularly appreciate the comprehensive Methods section which highlights the complexities involved in ML training and gives critical details on reproducing results and applying these approaches to similar scenarios. The justification for choosing both ML methods (L107) is also very useful to an ML novice. While not an ML expert albeit as someone with a statistical background, the decisions made by the authors as documented in Sections 2-4 intuitively make sense. The results are also intuitive – the models show skill over stations with long observation datasets / close proximity to other stations, and reduced skill elsewhere. I have some minor comments which I think would improve the manuscript but otherwise I think the paper is an excellent fit for the journal.

General comments

G1: The greatest source of uncertainty I think is in the spatial distribution of the surface observations (Fig. 1) which is very Europe-centric. Additionally, all/most of the surface sites are over land which diminishes the skill at predicting over oceans where the sources of aerosol are frankly very different to over land. I think the manuscript should be re-framed as a "Global land network of accumulation number" rather than global as there is limited evidence of skill over the oceans. I do not think this diminishes the paper but is more reflective of the results and limitations of the study. If the authors can provide some evaluation of concentrations over the ocean, even if rudimentary, that would be useful

> This comment is very accurate, and we certainly acknowledge the issue of strong variability in spatial representativeness of the utilised data sets. It is true that evaluating against observations over the oceans (and other underrepresented environments) would be useful in demonstrating clearly the current uncertainties. However, including more data at this stage would require a thorough consideration of which data is good for the comparison and which data set we should utilise and which not, would result in a considerable amount of extra work, and would also add further material to an already

quite lengthy manuscript. However, in the future we aim to extend the data in a follow-up manuscript where more data sets from marine and polar areas are included.

We also hesitate to change the title of the manuscript. While the accuracy of the modelled N100 fields vary between different environments, the fields themselves are global. We prefer to clarify the (expected) representativeness of the fields in the manuscript. Due to this excellent comment, we did realise that in the original version we did not clearly discuss the expected representativeness of our model in different environments. We now mention the underrepresented sections more specifically in the abstract (line 8 in the revised manuscript):

Instead of "However, performance declines in underrepresented regions and conditions, such as clean and remote environments, underscoring the need for more diverse observations." we now say: "However, performance declines in underrepresented regions and conditions, such as clean and remote environments including marine, tropical and polar regions, underscoring the need for more diverse observations."

We added the following paragraph to Sect. 5.4 (line 683 in the revised manuscript):

"[For the N100 measurements, the main challenge is data availability. To train ML models that capture diverse environments and meteorological conditions, we require a broad dataset that covers a wide range of locations and time periods.] In an ideal case, the dataset would represent environments with different natural and anthropogenic emission levels extending from low to high global extremes, as well as a wide spectrum of different anthropogenic to natural contribution ratios. The global distribution of long-term data sets, reflected by measurement stations utilised in this study, is clearly biased towards continental, anthropogenically influenced and European environments. Thus, the performance of our global ML models is expected to be worse in marine and tropical environments, as well as in the southern hemisphere and in polar regions. In addition, i[deally, we would have at least five years of data from each station.]"

We also replaced the following text from the last paragraph of the manuscript (line 780 onwards in the revised manuscript)

"Further evaluating the performance and reliability of the global MLR and XGB models in different environments and conditions will require additional data. We hope future collaborations will provide access to a wider measurement dataset, including data from stations not currently included in this analysis and more data from stations already part of the study. Although adding new data from measurement stations does not provide a global reliability estimate, it will allow us to assess model performance in new environments and conditions with unseen data."

with

"Improving and better evaluating the performance and reliability of the global MLR and XGB models in different environments and conditions will require additional data. We hope future research investments and collaborations will provide access to a wider long-term measurement dataset, extending especially towards marine, tropical, southern hemisphere and polar areas that are underrepresented in the current study. Although adding new data from such measurement stations does not provide a global reliability estimate, it will allow us to improve and assess the model performance in new

environments and conditions with unseen data. Including longer data sets from stations already part of this study will also improve the models, due to capturing more variability in the atmospheric conditions at these sites."

G2: The CAMS fields used as predictor variables (Table 2) are predominantly gas and aerosol tracers in that model, with some limited data assimilation of aerosol (from satellite AOD) but as far as I'm aware not gases. The paper lacks a quantification of the uncertainty in these predictor variables at the surface sites and in general. This is not to say that the use of these predictor variables is wrong, but I would appreciate some quantification of the relative uncertainty in these variables. If the surface sites measured the variables (e.g., EMEP over Europe) that would be a useful way to evaluate this uncertainty. Some qualitative evaluation is provided in Sections 5.3 and 5.4 but this should be extended.

It is certainly correct that the uncertainties in CAMS variables cause uncertainty in our ML models. It is also correct that an appropriate investigation of CAMS biases at the measurement stations would likely quantify these uncertainties. However, detailed quantification of uncertainties across the different CAMS variables is beyond the scope of the current manuscript. The evaluation of uncertainties related to CAMS aerosol variables are discussed thoroughly in Block et al. (2024), as mentioned in our manuscript, and those related to CO and NO2 have been presented by Inness et al. (2019) and Langerock et al. (2024), which we mention more clearly in the revised manuscript (see below). Based on Langerock et al. (2024), both of these gas concentrations are assimilated against satellite observations, and as the aerosol products are assimilated against AOD determined with satellite retrievals, the uncertainties in these variables can presumably be determined with much less regional variability than the variables that are measured and assimilated with worse and more heterogenous spatial coverage. Due to the extensive investigation required for quantifying the uncertainties in the applied variables, we prefer not to do that but express the related uncertainties more clearly in the manuscript.

Throughout the manuscript, we checked the use of word "bias" and changed it to word "uncertainty" where necessary.

We added the following in the Methods-section (line 229 in the revised manuscript):

"We discuss the possible effects of these and other CAMS variable related uncertainties in Sections 5.3 and 5.4"

We extended section 5.4 (starting at line 702 in the revised manuscript) from

"Regarding reanalysis data, challenges stem from various biases inherent in the CAMS and ERA5 reanalysis datasets. Block et al. (2024) provide a comprehensive discussion of the biases affecting CAMS aerosol variables, including uncertainties in polar regions due to limited satellite retrievals, omissions such as volcanic activity, and specific volcano-related biases around locations like Mauna Loa (Hawaii, USA) and Altzomoni (Mexico) —both of which appear as hotspots in our MLR model results. CAMS also does not model nitrate aerosol mixing ratios and represents hydrophilic and hydrophobic BC and OM mixing ratios with simplified partitioning based on emission fractions and a conversion rate over time (see Block et al. (2024) and the references therein). Although

we did not explore these biases for CAMS gas concentrations and meteorological variables, similar issues are likely present, potentially introducing some errors into our global N100 fields."

to

"Regarding reanalysis data, CAMS and ERA5 are subject to various uncertainties that can affect the performance of our ML models. Block et al. (2024) provide a detailed overview of uncertainties in CAMS aerosol variables, including limited satellite retrievals in polar regions, omissions such as volcanic activity, and specific volcano-related biases around sites like Mauna Loa (Hawaii, USA) and Altzomoni (Mexico)—both of which emerge as hotspots in our MLR model results. Additionally, CAMS currently excludes nitrate aerosol mixing ratios (Inness et al., 2019) and applies a simplified partitioning scheme for hydrophilic and hydrophobic BC and OM based on emission fractions and a time-dependent conversion rate (Remy et al., 2022). We should also note that the relations between N100 and OM or BC in our ML models are likely to be affected by the apparent challenges by CAMS in predicting the overall concentration levels of OM (Amarillo et al., 2014) or past changes in BC concentrations over areas such as China (Li et al., 2024).

For gas compounds, CAMS variables assimilated with satellite retrievals—such as CO and NOx—have been evaluated in studies by Inness et al. (2019) and Langerock et al. (2024). In contrast, variables not assimilated with satellite data are less thoroughly investigated, and their uncertainties likely vary notably across variables and regions. Although we do not explicitly assess the impact of these CAMS uncertainties on our ML model, they are expected to introduce errors into our global N100 fields."

To the last paragraph of section 5.4, when discussing issues stemming from sub-grid scale variability (line 719 in the revised manuscript):

"[This discrepancy], along with other CAMS uncertainties, [may partly explain the poor performance observed at some stations, even when using single-station models.]"

G3: The Brock et al reference seems like an interesting counterpoint, and I would have appreciated a direct comparison of the results of the 2 different approaches to deriving aerosol number concentrations, even if only over the measurement sites.

There are two main reasons why we did not do a direct comparison to Block et al., 2024. Firstly, they produced estimates of the activated CCN concentrations whereas we produced estimates of aerosol particles larger than 100 nm. While these estimates correlate, especially if looking at CCN at 0.4% supersaturation, these are not exactly the same. Secondly, Block et al., 2024 and the dataset they produced was published after we had already finalised our analysis, and we concluded that the comparison would be significant amount of additional work at that stage.

Qualitatively we can say that the comparison between observed and estimated CCN in Block et al., 2024 (Figure 7a) yields quite similar results to our comparison between observed and estimated N100 (Figure 5). Most of the estimated values in both cases remain within a factor of 10 from the observations and captures the range of

magnitudes. However, it should be noted here that our comparison between observed and estimated N100 is done using in-situ ground measurements and includes mainly continental measurement sites, whereas the comparison presented in Block et al is between their estimated CCN values and observed atmospheric radiation measurement (ARM) CCN near the surface and they include both continental and marine sites in the comparison.

Block et al. (2024) supplementary material contains also a global map of estimated CCN at 0.4% supersaturation in the lowermost 1 km for 2003-2021 (Figure C1). Comparing this to our global maps of N100 values for 2013, we see that they show similar features. The concentrations are low in polar regions and high in parts of South Asia and East Asia. However, there are also differences, for example our estimates have more variation in concentrations over the continents. Additionally, our estimates have much lower concentrations over oceans. It is likely that the marine concentrations are better estimated by Block et al., as our dataset had very limited representation of marine environments and Block et al. show quite good performance also in the marine environments (Fig 7a).

Specific comments

S1: [L36] "need to be captured within a factor of 1.5 of their true values" – this is mentioned tangentially in Rosenfeld et al. without context. Is there any reasoning behind the factor of 1.5? If so, please include

There is no specific reasoning for this factor, but we consider it as an expert estimate by Rosenfled et al. We have modified the respective sentence to reflect this a bit better:

"For example, Rosenfeld et al. (2014) estimate that global CCN concentrations should be captured within a factor of 1.5 of their true values to reliably assess aerosol effects on clouds."

S2: [L41] This is optional but I wonder if you can mention any specific regions/seasons which are blighted by lack of CCN observation coverage from e.g., too much cloud cover, lack of satellite data, etc

We think this further explanation is not required, as the main problem of CCN satellite observations is that the number concentrations cannot be reliably determined from satellite data, despite the cloud coverage conditions or spatial data coverage. We think this is explained clearly enough in the manuscript.

S3: [L125] What is the difference between testing/validation and holdout data? As an ML novice, this jargon appears to describe very similar partitions of the data phase space and might warrant a one-line explanation as to how they differ

Testing set, validation set, and holdout set are machine learning terms that are used somewhat interchangeably in the literature, and the concepts are very similar. However, the holdout set differs from the other two as it explicitly refers to the special case where the data is never used in any capacity to develop, tune or train the machine learning models but is held out until testing the very final version of the models. To follow the suggestion and make this clearer, we changed lines 126-128 in the revised manuscript from

"Testing the final version of the ML model is commonly done with a holdout set. The holdout set is separated from the training set at the beginning of the analysis and is reserved solely for testing the final version."

to

"To maintain the independence of the datasets used for training and testing the model, the model performance of the final ML model is typically evaluated with a dedicated dataset called the holdout set. This subset of the full dataset is set aside from the training data at the beginning of the analysis and reserved solely for testing the final ML model at the end of the analysis."

S4: [L165] Are the thresholds for 'excellent', 'good', and 'poor' fit at 0.2/0.3 arbitrary or established in the ML sciences? I understand the need for categorising model skill but perhaps there should be a neutral level between good and poor.

Thank you for the comment. The thresholds are indeed arbitrary. We needed to categorise model skill, so we defined these thresholds by looking at the model performance evaluation scatterplots. We have now changed these limits to 'good' (<0.2), 'adequate' (0.2<x<0.3), 'poor' (>0.3) throughout the paper.

S5: [L208] Just a quick check that the altitude of the interpolated predictor variables matched the altitude of the measurement station and corresponding CCN?

We did not do any adjustment of the predictor variables from CAMS / ERA5 to account for potential differences between station elevation and the elevation in the reanalysis. This is because most of the stations are at relatively low elevations and are in areas with relatively homogenous terrain. However, we have now checked the difference between the actual station elevation and the reanalysis elevation and for the majority of stations this is small (median height difference was 43m). There are a few notable exceptions to this (e.g., Mukteshwar (India), Schauinsland (Germany), Hohenpeissenberg (Germany) and Amman (Jordan), where the differences range from 308 m to 1496 m). However, we do not think this has an impact on the conclusions of this study.

[Section 4] This is a very useful section and I thank the authors for their level of detail

S6: [L503] Presumably the HAD issue is dust or is it NPF related - ACP - New particle formation, growth and apparent shrinkage at a rural background site in western Saudi Arabia?

We do not see why the dust or NPF should be captured less efficiently in HAD than in other locations. Our data set still contains sites as UAE (United Arab Emirates) and AMM (Amman, Jordan) where high dust concentrations are common and NPF is frequent in many of the sites. However, we think that a possible reason for discrepancies in HAD might be related to strong emission and concentration gradients near the site (oil refineries and vegetation around Yeddah, surrounded by sea and desert). Thus, the relatively coarse spatial resolution of the CAMS data may not reflect well enough the conditions at the site.

We have added to line 518 in the revised manuscript the following text:

"[In Hada al Sham, Saudi Arabia (HAD), both station-excluded models underestimated N100, whereas among the station-included models, the MLR model showed some improvement and the XGB model improved noticeably (Fig. 7b)]. The underestimation

likely stems from the station's complex surroundings, which includes desert, sea, and a nearby hotspot of anthropogenic and biogenic activity (Hakala et al., 2019). While actual concentrations at the station can be high due to the hotspot, reanalysis data cannot resolve such sub-grid scale variability, resulting in underestimated predictor values and low N100 estimates. The station-included XGB model may still perform well if the predictors maintain a correlation with N100, even when underestimated."

In section 5.4 we continued the discussion about challenges in station-included models on line 653 in the revised manuscript:

"[Since the models—especially the MLR model—struggle to capture conflicting variable-N100 relationships, stations with unique interactions relative to the rest of the dataset tend to experience the largest performance decline from single-station models to station-included models.] The unique interactions that lead to performance decline may arise not only between observed N100 and real aerosol, gas, or meteorological variables, but also artificially between observed N100 and CAMS variables distorted by uncertainties. One example of such artificial interaction is the sub-grid scale variability–related underestimation seen in MLR estimates at Hada al Sham, discussed in Sect. 5.1.3."

S7: [Section 5.1] This section is currently a little weak and would benefit from a hypotheses over why we see certain biases (see comment above). Potentially this would be useful also to the CAMS model developers. The line in [L515] starts to do this.

It is true that Section 5.1 does not originally address the reasons behind the challenges in ML model performance, as this discussion was primarily concentrated in Sections 5.3–5.4. However, in response to this suggestion and the earlier comment (S6), we have now slightly expanded the discussion for the time series at each station in Section 5.1.3. This ensures that all stations include some discussion of model behaviour. In addition to the changes made based on the previous suggestion, we have also added a clarification on line 528 in the revised manuscript.

"[In Värriö, Finland (VAR), the models performed well during summer, but the station-excluded models overestimated the low concentrations during winter (Fig. 7d).] While the MLR station-included model did not yield notably better results than station-excluded models, the XGB station-included model successfully captured the winter periods as well. In general, low concentrations tend to be quite difficult for our ML models to capture, but the station-included XGB model likely succeeds in capturing them because tree-structure allows it to fit more closely to any included training data.

S8: [L521] Potentially my only suggestion for adding to the methodology here which I think is optional is: did you try to re-train the ML model without one of either BC or OC given their significant correlation? Potentially there is some important detail which is missed by ignoring these predictor variables which could be useful if only including one?

The suggestion is very good, especially from ML point of view. Utilising strongly correlating variables as predictors is considered redundant and may cause overfitting. We did notice this issue and investigated it in quite much detail. With a previous version of the model, a few years ago, we investigated leaving only BC or OM, but it worsened the result, so we did not continue with it. The decision for continuing with both was also supported by BC and OM being partly related to very different aerosol sources, as OM

includes also secondary organic aerosol (SOA), formed in the atmosphere after photo-oxidation of both biogenic and anthropogenic volatile organic compounds (more in the answer for question M16 by Referee 2). While the representation of secondary organic aerosol is uncertain in global models (also reflected by such a strong correlation between BC and OM in CAMS, suggesting CAMS OM is dominated with primary organic matter), it can be expected to improve in the near future, due to significant investments in better understanding SOA formation. Due to these reasons, we decided to keep both BC and OM included. We still tested how the models performed when considering the ratio of BC and OM mixing ratios, but since it did not change the results dramatically (i.e., in terms of other important variables), we decided not to show these results, as they would have required still more model runs and added another complicated side path to the manuscript, without clear benefits for the outcome.

Thus, our decision was to include all parameters despite their role in the "N100-formation chain", e.g., including temperature, monoterpene concentrations and OM, temperature being the driver for monoterpene emissions and monoterpenes forming OM after atmospheric oxidation. This decision was supported by our tests with adjusted R-squared suggesting the model performance was not artificially improved due to overfitting issues (Sect 4.3.3).

To better justify the roles of including the different predictor variables, we have modified the following paragraph (starting at line 213 in the revised manuscript):

"From CAMS and ERA5 datasets, we selected reanalysis variables known to either directly or indirectly influence the formation, growth, losses or dilution of aerosol particles. Most variables were sourced from the CAMS dataset, while boundary layer height (BLH) was obtained from the ERA5 dataset. The list of reanalysis variables used as predictors is provided in Table 2."

To be like this:

"The list of reanalysis variables used as predictors is provided in Table 2. Most variables were sourced from the CAMS dataset, while boundary layer height (BLH) was obtained from the ERA5 dataset. The selected reanalysis variables are known to influence N100 either directly or indirectly. The variables with direct influence relate to primary emissions in the N100 size range (black carbon, organic matter in terms of primary organic matter, sulphate aerosol, smallest size ranges of dust and sea salt aerosol) and their sinks (rain). The variables with indirect influence either contribute to secondary aerosol formation and thus particle growth into N100 size range (sulphur dioxide, ammonia nitrogen monoxide/dioxide, terpenes, isoprene, organic matter in terms of secondary organic aerosol, temperature, relative humidity), or affect their transportation and dilution, or indicate general exposure to combustion and biomass burning in the air masses (wind speed, BLH, carbon monoxide). Many of these variables can be related to multiple processes affecting N100 concentrations, as discussed in Sect 5.2.1."

S9: [L534] Presumably the CO mixing ratio importance is because it's a useful proxy for biomass burning smoke? It seems like an odd fit here given its limited aerosol chemistry so would be good to identify a reason for its inclusion

It is true that CO mixing ratio is an odd one in the final list of variables. It is a very good tracer of combustion related primary particles, including the biomass burning smoke, as

suggested by the referee. This seems to show in the relatively high importance of CO in both models (Fig. 8 of the manuscript), especially in the MLR model. This is now expressed in the manuscript as shown in the answer to question S8 above.

S10: [L545] This relates to my General Comment – I think that the model framework with no marine sites will have very limited skill over the oceans. Perhaps the ocean ML-predicted concentrations can be tested with satellite, field campaigns or ship measurements where available but the negative coefficient for sea-salt is highly suggestive

It is true that without marine sites the ML model framework is likely to have limited skill over marine regions, and this is an area that requires further development. We are currently working on a follow-up study that incorporates newly available data from stations with marine influence, including two located in the Atlantic Ocean. These additions are expected to improve the model's performance over the oceans.

At this stage, we are not planning to use field campaigns or ship-based measurements. While such data could, in principle, be used for model evaluation, they typically lack the temporal coverage needed to train ML models effectively. Our preference is to use datasets that span full seasonal cycles to ensure robust learning. Ship measurements pose additional challenges, particularly in terms of spatial collocation with reanalysis data, which is more straightforward with fixed measurement stations.

Regarding satellite data, we have not included them because they do not directly measure aerosol number concentrations. Instead, they infer aerosol loading based on radiative properties, which cannot be reliably translated to number concentrations around 100 nm (e.g., Rosenfeld et al., 2014). Moreover, satellite observations generally represent the entire atmospheric column, whereas our focus is on near-surface concentrations.

Block, K., Haghighatnasab, M., Partridge, D. G., Stier, P., and Quaas, J.: Cloud condensation nuclei concentrations derived from the CAMS reanalysis, Earth System Science Data, 16, 443–470, https://doi.org/10.5194/essd-16-443-2024, 2024.

Hakala, S., Alghamdi, M. A., Paasonen, P., Vakkari, V., Khoder, M. I., Neitola, K., Dada, L., Abdelmaksoud, A. S., Al-Jeelani, H., Shabbaj, I. I., Almehmadi, F. M., Sundström, A. M., Lihavainen, H., Kerminen, V. M., Kontkanen, J., Kulmala, M., Hussein, T., & Hyvärinen, A. P.: New particle formation, growth and apparent shrinkage at a rural background site in western Saudi Arabia. Atmospheric Chemistry and Physics, 19(16), 10537–10555. https://doi.org/10.5194/acp-19-10537-2019, 2019

Inness, A., Ades, M., Agusti-Panareda, A., Barre, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M., and Suttie, M.: The CAMS reanalysis of atmospheric composition, Atmospheric chemistry and physics, 19, 3515–3556, https://doi.org/10.5194/acp-19-3515-2019, 2019.

Langerock, B., Arola, A., Benedictow, A., Bennouna, Y., Blake, L., Bouarar, I., Cuevas, E., Errera, Q., Eskes, H. J., Griesfeller, J., Ilic, L., Kapsomenakis, J., Y Li, C. W., Mortier, A., Pison, I., Pitkänen, M., Richter, A., Schoenhardt, A., Schulz, M., ... Zerefos, C.; Validation report for the

CAMS global reanalyses of aerosol and reactive trace gases, years 2003-2023. ECMWF Copernicus Report. https://doi.org/10.24380/g8h7-kd21, 2024

Rosenfeld, D., Andreae, M. O., Asmi, A., Chin, M., Leeuw, G., Donovan, D. P., Kahn, R., Kinne, S., Kivekäs, N., Kulmala, M., Lau, W., Schmidt, K. S., Suni, T., Wagner, T., Wild, M., and Quaas, J.: Global observations of aerosol-cloud-precipitation-climate interactions, Rev. Geophys., 52, 750–808, https://doi.org/https://doi.org/10.1002/2013RG000441, 2014.

Citation: https://doi.org/10.5194/ar-2025-18-RC1

RC2

Review of "Global fields of daily accumulation-mode particle number concentrations using in situ observations, reanalysis data and machine learning" by Aino Ovaska and co-authors.

In this manuscript the authors use a database of N100 measurements from 35 stations combined with reanalysis data (ERA5 meteorology and CAMS aerosol) to train and test two established machine learning (ML) models. N100 is a good proxy for CCN concentrations, which itself is an important for cloud microphysical and radiative properties. There is paucity of in-situ measurements for both variables, which results in considerable uncertainty when evaluating climate models and estimating future projections. A ML model that can provide robust global estimates of N100 and CCN concentrations would be a very important step forward for the community, therefore the focus of this study is very relevant.

The authors take commendable effort towards robustly training and testing the ML models. All steps are considered and justified throughout the manuscript. The results are well presented and discussed and framed very well with regards to associated limitations and steps required to improve on the ML models accuracy and representativeness.

I thoroughly enjoyed reading this manuscript and recommend publication in Aerosol Research following a discussion on a few largely minor comments.

General comments

G1: Some regions may have meteorological drivers that are not found in other regions – for example the Southern Ocean. The synoptic / seasonal meteorology and sources of aerosol will be very different here than any other region that includes your training dataset. The time series analysis in Figure 7 (lines 409 – 502) for Alert (located in a relatively remote region) demonstrates that it struggles with this. Does this limit the use of the ML models in regions very far from any of the stations? (e.g., a lot of the southern hemisphere). I should note that this is still an excellent dataset and just clearly demonstrates the need for additional measurements in these remote regions.

This is correct and we touch on this in sections 5.2.2. and 5.3. Regions with unique meteorological conditions or aerosol sources pose a challenge for our ML models, because these conditions are not represented in the training dataset. If we exclude Alert from training, the model struggles there and likely in the surrounding Arctic area, which is both remote and meteorologically distinct. Similarly, our training set does not contain marine measurement stations, and therefore the ML models likely struggle over the ocean. This limitation is not necessarily solely due to physical distance from stations, but rather due to differences in meteorology, emissions, or other region-specific factors.

Interestingly, in regions like the east coast of Australia, which is geographically distant from our measurement stations, the two ML models produce consistent results. This suggests that the region's conditions are sufficiently represented in the training data by the measurement stations, despite the distance.

We are currently working on a follow-up study that expands the training dataset to include more stations (from which data has only just become available to us), including some with marine influence. However, even with these additions, certain regions, such as the Southern Ocean, will remain underrepresented. Without measurements from these areas, the ML model's N100 estimates are likely to be less reliable in these regions.

G2: This is a comment rather than a suggestion. Given the lack of observations in many regions – I wonder whether one way to test this is to repeat the methods using output from a global aerosol microphysics model. There would still be associated uncertainty due to the microphysical processes but it would be a very good test of the methods.

If we understand correctly, the referee is suggesting we use output from a global aerosol microphysics model such as GLOMAP instead of observations of N100 as the target (predicted variable). This is an interesting idea, and a few previous studies have attempted this (e.g. Yu et al., 2022; Li et a., 2022). However, as the reviewer highlights, the ML models will inherit any bias from the physics-based aerosol microphysics model. As this additional analysis would be a vast amount of work, and would considerably lengthen the manuscript, we have decided not to pursue this.

G3: How would you expect the ML models to perform in a PI scenario where you have removed the most important features (the anthropogenic aerosol sources?). Do you think there is a possibility that they are only representative of the PD environment?

Our ML models are only representative of the PD environment. ML models can replicate only conditions they are trained on, which here spans 2003-2020. The models do not understand the real underlying physics and chemistry of aerosol formation and growth but estimate it based on the connections between our predictor and target variables. If some of these connections are drastically different from what the ML models are trained on, such as in PI or far future situations, the models cannot be trusted to produce physically reasonable results. Therefore, for example using low values for anthropogenic predictors would not replicate PI N100.

We now underlined this by adding the following to the conclusions (line 777 in the revised manuscript):

However, it should be noted that ML models trained with observational data as the target variable cannot be expected to represent these variables reliably in too distinct conditions – determining pre-industrial or future N100 cannot be done based on present day observations.

G4: Overall, do you think the community can use this as a realistic proxy for N100 (outside of Europe – Line 647) in the absence of a detailed aerosol microphysics model? Is there sufficient skill? Or do you believe more work is required?

We are confident that already the current version of the N100 dataset can be considered realistic, especially in continental areas. It should be noted that the training data for the final global models include all the acquired datasets, thus including training data for those environments that were considered underrepresented due to cross-validation results (showing clear improvement from station-excluded to station-included models).

For the environments unrepresented in the training data, our results should be considered with care. However, more data and updated ML models trained on that data will certainly improve the results. Currently we are gathering a wider global data set, including data from marine and polar environments, as well as continental environments influenced by marine air masses, that was not readily available before. In the answer for question G4 for Referee 1, we present more detailed descriptions of the training data that is expected to improve the global representativity of the models and the related modifications made to the manuscript.

Minor comments

M1: Line 194. The aerosols will likely have diurnal cycles in some locations. Therefore, there is an implicit assumption that the model reanalysis is able to capture the diurnal cycle correctly – is this a valid assumption? Why not use 6hourly? Some predictors (e.g., u, v, T, RH, some aerosol emissions) will likely vary throughout the diurnal cycle – and may be overlooked during feature importance analysis etc.

This is an important point. Indeed, many of the predictors exhibit diurnal variability in the real atmosphere, and it is theoretically possible to extend our method to sub-daily timescales. However, for the purposes of this study, we chose not to pursue that direction. While CAMS provides 3-hourly data and ERA5 offers hourly resolution, we are cautious about assuming that reanalysis products fully capture these diurnal cycles with sufficient accuracy. Furthermore, capturing diurnal variability in N100 using ML models would introduce additional complexity and variance to an already challenging prediction task. For this reason, we chose to focus on daily averages in this study.

There are also practical considerations: moving to 3-hourly or 6-hourly resolution would increase the dataset size by a factor of 4–8, requiring significantly more computational resources. It would also necessitate careful handling of local solar time, rather than the UTC-based approach we currently use.

That said, we agree that daily averaging may reduce the apparent importance of variables like boundary layer height (BLH), which have strong diurnal cycles. If sub-daily

predictions were pursued in future work, we expect that the feature importance of such variables could increase.

M2: Line 196. How many data points did this remove from the sets?

Outlier removal removed 179 datapoints out of the original 49 669 days in the training set, leaving the 49 490 days mentioned on line 196.

M3: Line 197. As you often measured very low concentrations did you also include zero counts when calculating the daily mean?

We did not remove zero values before calculating the daily N100 means. However, zero values were present only at two stations, and in total, just 15 daily means included any zero values. After computing the daily means, we applied a $\log_{10}$ transformation. To enable this, all N100 values less than or equal to one were replaced with the smallest N100 value above one measured at the respective station.

M4: Line 212-217. How were gridded datasets spatially collocated with the measurement stations? Was the grid cell average used or linearly interpolated from nearby neighbouring grid points?

We used values interpolated to the coordinates of the measurement station.

We altered following places in the manuscript:

Line 222 in the revised manuscript from

"We matched the reanalysis data to the N100 measurements by selecting data from the grid-cells containing the measurement stations and including only days with observations."

to

"In these sets, we collocated the reanalysis data to the N100 measurements by using values interpolated to the point of the measurement station and including only days with observations."

Line 225 (revised manuscript) from

"However, it should be noted that average conditions within a grid-cell (up to 80 km) may sometimes fail to represent the single-point measurements due to sub-grid scale variability in emission sources, meteorology, and topography."

to

"However, the average conditions within a grid-cell (up to around 80 km) and even the interpolated values may sometimes fail to represent the single-point measurements due to sub-grid scale variability in emission sources, meteorology, and topography."

Line 230 (revised manuscript) from

"In addition to the training and holdout sets, we used reanalysis data as input for generating the global N100 fields, retrieving the reanalysis data covering the whole globe for 2013."

to

"In addition to the training and holdout sets, we used reanalysis data as input for generating the global N100 fields, retrieving the reanalysis data covering the whole globe at 0.75x0.75-degree resolution for 2013."

Line 233 (revised manuscript) from

"We first adjusted the 0.25x0.25-degree resolution of the ERA5 dataset to match the 0.75x0.75-degree resolution of the CAMS dataset by calculating grid-cell averages that correspond to CAMS data grid size."

to

"For the global fields, we first adjusted the 0.25x0.25-degree resolution of the ERA5 dataset to match the 0.75x0.75-degree resolution of the CAMS dataset by calculating grid-cell averages that correspond to CAMS data grid size. The rest of the analysis proceeded the same for all sets."

Line 717 (revised manuscript) from

"Because reanalysis data represents grid-cell averages, it may not capture the true predictor variable concentrations at the measurement site, leading to biases in the model's learned relationships."

to

"Because reanalysis data represents grid-cell averages, it may not capture the true predictor variable concentrations at the measurement site, even if the reanalysis data is interpolated to the exact station location, leading to uncertainties in the model's learned relationships."

M: Line 221. Related to above, how were gridded datasets collocated with the altitude of the measurement stations? Stations on a hilltop or a mountain site may not be well represented by the model surface mean.

We did not do any adjustment of the predictor variables from CAMS / ERA5 to account for potential differences between station elevation and the elevation in the reanalysis. This is because most of the stations are at relatively low elevations and are in areas with relatively homogenous terrain. However, we have now checked the difference between the actual station elevation and the reanalysis elevation and for the majority of stations this is small (median height difference was 43m). There are a few notable exceptions to this (e.g., Mukteshwar (India), Schauinsland (Germany), Hohenpeissenberg (Germany) and Amman (Jordan), where the differences range from 308 m to 1496 m). However, we do not think this has an impact on the conclusions of this study.

M7: Line 312 Downweighing = Downweighting?

Thank you for the comment. Changed all instances of downweighing to downweighting.

M8: Line 336. Re manually selecting parameter combinations. Was there not a statistical method that could be used to eliminate any human sourced bias?

We selected the hyperparameters using grid search, which is a brute-force method where each hyperparameter is given a range of values and search iterates over all the possible combinations. The user chooses the initial hyperparameter ranges to test and manually selects the final hyperparameters, so there is a level of subjectivity involved. This is a commonly used and simple method, and we considered it to be sufficient for our use. There are more complex statistical methods, but these can be more challenging to implement and do not necessarily yield better results. Different methods are also likely to produce different optimal hyperparameter combinations so the selected method will affect the final selected hyperparameters.

To clarify this, we edited the text starting on line 158 in the revised manuscript from

"To optimize the HPs, we employed CV, where different combinations of HPs were evaluated to identify the combination of hyperparameters that yielded the best performance across several validation folds."

to

"To optimize the HPs, we employed grid search, which is a commonly used brute-force method where each hyperparameter is given a range of manually selected values and the search iterates over all possible combinations. The search can be repeated multiple times focusing only on a subset of hyperparameters or using narrowing ranges of values based on previous rounds. We evaluated the performance of each hyperparameter combination using CV and selected the combinations that yielded the best average performance over the validation folds."

and line 347 (revised manuscript) from

"We used the spatial train-validation split method for cross-validation to ensure the tuned HPs generalized across all stations (Table 3). One station, Schauinsland, Germany (SCH), was excluded from HP tuning due to its frequent positioning above the boundary layer during winter (Birmili et al., 2016). Based on the results, we manually selected parameter combinations that produced strong average RMSElog10 across cross-validation rounds."

to

"We used grid search and the spatial train-validation split method for cross-validation to ensure the tuned HPs generalized across all stations (Table 3). One station, Schauinsland, Germany (SCH), was excluded from HP tuning due to its frequent positioning above the boundary layer during winter (Birmili et al., 2016). Based on the grid search results, we selected parameter combinations that produced strong average RMSElog10 across cross-validation rounds."

M9: Line 367. Did you pay any specific attention in the analysis to how the ML models tested as a function of how remote the excluded station was?

We find that this topic is discussed in Sect. 5.1.2 (lines 481-503 in the revised manuscript).

M10: Line 406. weighed = weighted?

Corrected weighed to weighted.

M11: Figure 3. Suggest making figure wider to clearly show the notches and features of the boxplots

> Thank you for the comment. We made Figure 3 and the boxes in the figure wider for easier readability.

M12: Line 441 (and 213). 2020 to 2022 had a reasonably strong -ve ENSO index. Could this bias the comparison between training and testing RMSE values?

> We trained the models using data from before 2020, primarily around 2013. As a result, the models are likely better at capturing conditions similar to those in the training period. In principle, if the testing conditions differ—for example due to ENSO—the model's performance may degrade. This helps explain why testing RMSE values are higher than training RMSE values: the training data does not fully represent the variability present in the testing period. This discrepancy is not necessarily a sign of bias, but rather an indication of the model's ability (or limitation) to generalize to unseen conditions. That said, we do not know the specific impact of ENSO on global N100 concentrations, nor can we say that the observed decrease in model performance in 2020-2022 period is related to ENSO.

M13: Figure 5. Suggest adding 'MLR model' and 'XGB models' to the top of panels (a) and (c) to make it automatically clear to the reader what is different.

> Thank you for the comment, we added 'MLR' and 'XGB' text to panels a and c in Figure 5.

M14: Line 459. Seems XGB is best unless extreme values. Would you therefore recommend using a combination of both? **XGB unless MLR predicts values < 25 or > 5000**?

> This is an insightful suggestion. In an earlier stage of the study, we considered creating a hybrid model that we would estimate to be better than any of the two separate models. To do such hybrid in an abrupt manner as suggested by the referee, might, however, lead to sudden changes in concentrations, if XGB happens to predict much less extreme values for the surrounding days or grid cells than MLR for the day and/or grid cell in question. Instead, we would suggest possibly including a third model and then blending the model outputs together somehow, possibly using weighted averages. However, we are not currently sure how to proceed with this in practice, so we decided not to discuss this in the manuscript.

M15: Line 468. Suggest authors add number of stations where XGB < MLR and vice versa.

> Edited text starting on line 481 (revised manuscript) from
>
> "In terms of the training error (Fig. 6), XGB had typically lower or equally good RMSElog10 values than MLR, indicating better performance, though there were also stations where MLR performed better."
>
> to
>
> "In terms of training error (Fig. 6), 25 out of 35 stations showed lower RMSElog10 values for XGB compared to MLR, indicating generally better performance. However, MLR achieved equally good or better performance in 10 stations."

M16: Line 521. Given the importance of BC and OM, how well are they represented in CAMS?

As described in detail by Remy et al. (2022) and summarized briefly by Eskes et al. (2024), fixed fractions of OM and BC emissions are assumed to be in hydrophobic form and to turn into hydrophilic forms over fixed e-folding times in CAMS. The secondary part of OM (SOA) is represented with dedicated tracers, distinguishing biogenic and anthropogenic origins, and coupled with tropospheric chemistry for their production.

Some work has been done to evaluate OM and BC predicted by CAMS. Such evaluations indicate a substantial overestimation of OM, up to a factor 3, as compared with global surface observations (Amarillo et al., 2024). For BC measured in China, CAMS appears to capture the overall concentration levels and their seasonal cycles relatively well, but to fail in predicting the observed decline in BC concentrations over the past couple of decades (Li et al., 2024).

We added the following text into the end of this paragraph (line 676 in the revised manuscript):

"We should also note that the relations between N100 and OM or BC in our ML models are likely to be affected by the apparent challenges by CAMS in predicting the overall concentration levels of OM (Amarillo et al., 2014) or past changes in BC concentrations over areas such as China (Li et al., 2024)."

M17: Line 531. What happens if you were to remove either BC or OM as one of the predictor variables?

A similar question (S8) by Referee 1 is addressed above.

M18: Figure 9. Missing M in label of panel (c). Suggest making station circles in (c) larger.

Thank you for the comment. We corrected LR to MLR in the colorbar of panel c. We also increased the size of the station circles.

M19: Line 602. Would you therefore recommend concentrating on regions with extreme N100 magnitudes to better train the global ML models?

Ideally, one should cover the whole N100 concentration level scale encountered in the atmosphere, as well as regions having very different natural and anthropogenic emissions contributing to N100, including different anthropogenic to natural contribution ratios. Concerning the N100 levels, it might be more beneficial to focus on low-concentration regions rather than high-concentrations ones, given that cloud droplet number concentrations, and eventually many cloud properties, tend to be most susceptible to aerosol over areas having low levels of CCN (e.g. Reutter et al., 2009; Liu et al., 2024). In practice, the choice of stations is severely limited by the availability of data. At present, for example, there is scarcity of data from the southern hemisphere, as well from many large-scale ecosystems such as tropical forests, savannah etc.

The regions, where additional long-term data sets are estimated to be most useful, are now better stated in the revised manuscript. A thorough description of the changes is given in the reply to the first general comment (G1) by Referee 1.

M20: Line 667. Worth noting that in well mixed boundary layers these ground-based ML models will likely provide representative values at cloud base.

We adjusted text (starting on line 696 in the revised manuscript) from

"Additionally, because our method relies on ground-level N100 measurements, our ML models can generate only ground-level N100 estimates. However, for many applications knowing the vertical profile of N100 would be important. For example, CCN concentrations are particularly important near or above the cloud base (Quaas et al., 2020), especially in cases where the cloud base is decoupled from surface conditions (Su et al., 2024)."

to

"Additionally, because our method relies on ground-level N100 measurements, the ML models can only produce ground-level N100 estimates and do not provide vertical profile information, which is needed for certain applications. For example, when studying aerosol–cloud interactions, CCN concentrations near or above the cloud base are particularly important (Quaas et al., 2020). While the ground-level aerosol concentrations represent the cloud-level concentrations in well-mixed boundary layers, where surface and cloud base conditions are coupled, they do not reflect cloud-level concentration under decoupled conditions (Su et al., 2024)."

Amarillo A. C. et al.: Validation of aerosol chemical composition and optical properties provided by Copernicus Atmosphere Monitoring Service (CAMS) using gound-based global data, Atmos. Environ., 334, 120683, https://doi.org/10.1016/j.atmosenv.2024.120683, 2024.

Birmili, W., Weinhold, K., Rasch, F., Sonntag, A., Sun, J., Merkel, M., Wiedensohler, A., Bastian, S., Schladitz, A., Löschau, G., Cyrys, J., Pitz, M., Gu, J., Kusch, T., Flentje, H., Quass, U., Kaminski, H., Kuhlbusch, T. A. J., Meinhardt, F., … Fiebig, M.: Long-term observations of tropospheric particle number size distributions and equivalent black carbon mass concentrations in the German Ultrafine Aerosol Network (GUAN). Earth System Science Data, 8(2), 355–382. https://doi.org/10.5194/essd-8-355-2016, 2016

Eskes H. et al.: Technical note: Evaluation of the Copernicus Atmosphere Monitoring Service Cy48R1 upgrade of June 2023, Atmos. Chem. Phys., 24, 9475-9514, https://doi.org/10.5194/acp-24-9475-2024, 2024.

Li, J., Hendricks, J., Righi, M., and Beer, C. G.: An aerosol classification scheme for global simulations using the K-means machine learning method, Geosci. Model Dev., 15, 509–533, https://doi.org/10.5194/gmd-15-509-2022, 2022.

Li W. et al.: Evaluation of MERRA-2 and CAMS reanalysis for black carbon aerosol in China, Environ. Poll., 342, 123182, https://doi.org/10.1016/j.envpol.2023.123182, 2024.

Liu, J. et al.: Cloud susceptibility to aerosols: Comparing cloud-appearance versus cloud-controlling factors regimes, J. Geophys. Res. Atmos., 129, e2024JD041216, https://doi.org/10.1029/2024JD041216, 2024.

Quaas, J., Arola, A., Cairns, B., Christensen, M., Deneke, H., Ekman, A. M. L., Feingold, G., Fridlind, A., Gryspeerdt, E., Hasekamp, O., Li, Z., Lipponen, A., Mülmenstädt, J., Nenes, A., Penner, J. E., Rosenfeld, D., Schrödner, R., Sinclair, K., Sourdeval, O., … Wendisch, M.:

Constraining the Twomey effect from satellite observations: Issues and perspectives. Atmospheric Chemistry and Physics, 20(23), 15079–15099. https://doi.org/10.5194/acp-20-15079-2020, 2020

Rémy, S. et al.: Description and evaluation of the tropospheric aerosol scheme in the Integrated Forecasting System (IFS-AER, cycle 47R1) of ECMWF, Geosci. Model Dev., 15, 4881–4912, https://doi.org/10.5194/gmd-15-4881-2022, 2022.

Reutter, R. et al.: Aerosol- and updraft-limited regimes of cloud droplet formation: influence of particle number, size and hygroscopicity on the activation of cloud condensation nuclei (CCN). Atmos. Chem. Phys., 9, 7067-7080, https://doi.org/10.5194/acp-9-7067-2009, 2009.

Su, T., Li, Z., Roldan Henao, N., Luan, Q., & Yu, F.: Constraining effects of aerosol-cloud interaction by accounting for coupling between cloud and land surface. In *Sci. Adv* (Vol. 10). https://www.science.org, 2024

Yu, F., Luo, G., Nair, A. A., Tsigaridis, K., & Bauer, S. E.: Use of Machine Learning to Reduce Uncertainties in Particle Number Concentration and Aerosol Indirect Radiative Forcing Predicted by Climate Models. Geophysical Research Letters, *49*(16). https://doi.org/10.1029/2022GL098551, 2022

Citation: https://doi.org/10.5194/ar-2025-18-RC2