

Author Reply to Editor and Reviewer Comments

Round 2 (Minor Revisions; Reviewer #2 remaining points)

Dear Dr. Ladino,

We thank you for the favourable round-1 evaluation and for forwarding Reviewer #2's remaining final comments together with three technical corrections. We have addressed each of these points in a new tracked-changes manuscript. In what follows, we reply to each remaining point in order, indicating the corresponding manuscript change by section or figure. The numbering R1–R6 refers to the substantive remaining points; T1–T3 to the technical corrections.

Reviewer #2 remaining points

R1: Abstract — “elusive links” cautionary message

Reviewer comment

“We agree with the referee that the primary scientific message—that even with more than 500 co-located variables at one of the most heavily instrumented stations globally, strong links to INP concentrations remain elusive—is a valuable cautionary result, and we have committed more fully to this framing in the revised abstract.” As you know, I agree that this is an important scientific message, but I cannot find it spelled out any more clearly in the abstract now. You've added that “results underscore the need for site-specific parameterizations”, but nothing was added about elusive links.

Response

We agree that our round-1 abstract revision did not in fact spell out the cautionary, “elusive links” framing that we had committed to in the round-1 reply. We have now added an explicit sentence at the end of the abstract that states the cautionary finding directly, naming both the scope of the data (more than 500 co-located, high-resolution variables at one of the world's most heavily instrumented atmospheric stations) and the conclusion (no monitored parameter is strongly, deterministically linked to INP concentrations).

Change in manuscript

Abstract (end): Added the sentence *“A core cautionary finding of this study is that, even with more than 500 co-located, high-resolution variables at one of the world's most heavily instrumented atmospheric stations, strong, deterministic links between INP concentrations and monitored parameters remain elusive.”*

R2: Fig. 2 caption — how the displayed variables were selected

Reviewer comment

Original comment “Line 144” from Reviewer 1 and “How was the subset...” from Reviewer 2: “Only half of the variables are listed in Figure 2”. Your answer addresses the selection going from Fig. 2 to Fig. 3, but I understand Reviewer 1’s question as asking why there are not 84 variables listed in Fig. 2, i.e. how was the selection for those variables that are shown in Fig. 2 done? In the answer to Reviewer 2 you state that these are the top-ranked variables, but that information is still missing from the caption.

Response

We thank the editor for the clarification — this is indeed the missing piece in the Fig. 2 caption. We have added a short clause to the caption stating explicitly that only the top-ranked variables (by median random-forest importance across the 50-seed sweep; see R3a below) are shown, for legibility. The 84-variable analysis set itself is unchanged.

Change in manuscript

Fig. 2 caption: added the clause “*Only the top-ranked variables (by median random-forest importance across 50 seeds) are shown for legibility*”, immediately preceding the existing listing of the six variables that are also examined in Fig. 3.

R3a: Random-seed sensitivity quantified as an uncertainty

Reviewer comment

I still struggle with your statements around the ML methodology. On the one hand, you say that random seeds were observed to lead to differences, that results were not robust, and that a comparison would be artificial. On the other hand, you do insist that results were qualitatively similar and go ahead with showing one of them. [...] Wouldn’t the differences due to random seeds count as some sort of uncertainty, so that showing them would be necessary (just as you wouldn’t show one day of measurements in your figure and say that other days deviate but are qualitatively the same – instead you’d add an uncertainty bar of variability)?

Response

We accept this critique fully. Treating the random-seed spread as a methodological uncertainty rather than a footnote is the correct framing, and the previous round-1 wording (“qualitatively similar”) was not adequate. We have re-run the random-forest analysis with 50 independent random seeds (seeds 1...50). The regenerated Fig. 2 now reports, for each retained variable, the per-variable *median* importance across the 50 seeds together with 5–95th percentile whiskers — i.e., a non-parametric uncertainty envelope over seed choice. This makes the seed sensitivity quantitatively visible alongside the central estimate.

The seed-sensitivity protocol (number of seeds, the seed-aggregation statistic, and the whisker definition) is now also stated explicitly in the Methods (Sect. 2.2 / Sect. 3.2). The implementation

is in the deposited analysis scripts (see R3b): a multi-seed random-forest sweep produces the seed-resolved importance scores, and a companion plotting script regenerates the Fig. 2 panel.

Change in manuscript

Sect. 3.2: replaced the round-1 “qualitatively similar” phrasing with the new wording: “*we ran the analysis with 50 independent random seeds (1..50). Figure 2 reports the per-variable median importance with 5–95th percentile whiskers across seeds.*” A new sentence in the same subsection documents the seed-sensitivity protocol. Fig. 2 itself was regenerated from the 50-seed output.

R3b: Code and software-version sharing

Reviewer comment

I appreciate that you added the software packages used, but in my understanding the AR/Copernicus data policy requires you to share the scripts, the versions of the software used, etc. (https://www.aerosol-research.net/policies/data_policy.html). I don’t think these standards should be any less strict for a “hypothesis-generating study”, especially in light of the ambiguous results from the different ML methodologies. The editor may wish to comment on the extent to which software sharing and documentation is necessary here.

Response

We agree. The analysis scripts (the R analysis notebook, the multi-seed random-forest sweep, the figure-generation script, and the dust-episode screening script), together with the captured `sessionInfo()` for the exact R version and package versions used and the harmonised 20-minute feature matrix used as input to the random-forest analysis, are deposited as a single Zenodo record at <https://doi.org/10.5281/zenodo.20367476>. The Code Availability section at the end of the manuscript points to this DOI.

We thank the reviewer for pressing on this; the deposit raises the documentation standard of the study to what the AR/Copernicus data policy requires, irrespective of the hypothesis-generating framing.

Change in manuscript

We populated the Code Availability section at the end of the manuscript with a paragraph pointing to the published Zenodo deposit at <https://doi.org/10.5281/zenodo.20367476>, listing the analysis notebook, the multi-seed random-forest sweep, the figure-generation script, and the dust-episode screening script, and noting that the full R version and package versions used to generate the results are captured in the archived session-info output.

R4: “Other data . . . upon request” clause

Reviewer comment

Similarly, your statement “Other data are available upon request from the corresponding authors.” appears questionable with regards to the Data policy quoted above, which states “If the data are not publicly accessible at the time of final publication, the data statement should describe where and when they will appear, and provide information on how readers can obtain the data until then.” What are these “other data” and why can’t you make them publicly accessible?

Response

The editor is right that the round-1 “upon request” wording was not compliant with the AR/Copernicus data policy. We have removed that clause and replaced it with an explicit enumeration of what the “other data” actually are and where each is deposited. Concretely, the derived analysis-ready datasets (the 84-variable feature matrix harmonised to a 20-minute time base, the multi-seed random-forest importance table underlying the revised Fig. 2, and the dust-episode screen output — see R6 below) are deposited together with the analysis scripts in the same Zenodo record (see R3b).

Change in manuscript

Data Availability (end of manuscript): replaced the “upon request” clause with the explicit text *“Derived analysis-ready datasets (the harmonised 20-minute feature matrix used as input to the random-forest analysis, the multi-seed random-forest importance table, and the dust-episode screen output) are deposited together with the analysis scripts at <https://doi.org/10.5281/zenodo.20367476>.”*

R5: Fig. 2 y-axis labels — SMEAR short names plus spelled-out long names with units

Reviewer comment

Fig. 2 caption: I understand that since you’ve been dealing with this data a lot, the translation from the six expanded variable names to the SMEAR abbreviations you use on the y-axis is clear to you, but it is not obvious for someone else (me). [...] please highlight these variables in the plot, or at least give the shortened variable name in the caption where you explain the long names for the six variables [...]. Also, I insist that in general the names on the y-axis are quite uninformative and would find spelled-out names that include units warranted (perhaps as a secondary y-axis to add to the SMEAR names?). At the very least, please add to the caption that the short names are from SMEAR and where one can look up the long names.

Response

We have regenerated Fig. 2 with the y-axis now displaying, on each tick, the SMEAR-II short name paired with its spelled-out long name and units. The pairing therefore appears directly on the plot, alongside the median-and-whisker bars from R3a, so a reader does not have to translate between caption and plot to interpret a variable. The caption additionally states explicitly that the short names are SMEAR-II metadata identifiers and points to the SMEAR variable registry where the full set can be looked up.

Change in manuscript

Fig. 2: y-axis labels regenerated to show paired SMEAR short name + spelled-out long name + units (figure asset replaced via the regenerated plot from the seed-aware script). Fig. 2 caption: added the clause “*Short names on the y-axis are SMEAR-II identifiers paired with their spelled-out long names and units; see <https://smear.avaa.csc.fi> for the full SMEAR variable registry.*”

R6: Identification of dust episodes during the PINC winter period

Reviewer comment

Identification of dust episodes: I understand that you cannot “definitively confirm” dust transport episodes as a source of INPs, but the most heavily equipped Hyytiälä site for sure has measurements that can indicate dust episodes in general? Couldn’t you at least analyse whether there were any dust episodes during your winter measurements?

Response

We accept this point. We have implemented an explicit screen for documented long-range dust transport events to Finland during the PINC measurement window (19 February–2 April 2018) by cross-referencing the systematic Finnish Meteorological Institute catalog of long-range dust transport events to Finland for the period 1980–2022 (Meinander et al., 2023). This catalog identifies 86 documented dust intrusions over the 43-year period (59 Saharan, 22 Aral-Caspian, 5 Middle East); we find no documented event overlapping the PINC measurement window. The nearest documented Saharan dust depositions over Finland are dated to 8–10 March 1991 (northern Finland) and 21–23 February 2021 (southern Finland), neither within our campaign. The FMI catalog does not explicitly enumerate Icelandic or Greenlandic glacial-outwash dust events, but no major high-latitude glacial-dust outbreak affecting Northern Europe in spring 2018 has been reported in the peer-reviewed literature to our knowledge.

We therefore find no documented evidence that long-range dust transport materially contributed to the winter INP signal observed by PINC. This negative finding is added to Sect. 3.3.1 of the revised manuscript and resolves the round-1 hedge about glacially sourced dust transport not being excludable a priori. As an additional cross-check, the dust-screening script and a populated candidate-date list are deposited with the analysis code; the script computes the aethalometer absorption Ångström exponent from SmartSMEAR aethalometer channels on any candidate date that an interested reader may wish to inspect.

Change in manuscript

A new paragraph is added at the start of Sect. 3.3.1 (Previous parameterizations) reporting the negative outcome of the FMI catalog screen, citing Meinander et al. (2023), and connecting the finding to the choice of the dust-independent DeMott et al. (2010) formulation over the dust-aware DeMott et al. (2015) revision. The Meinander et al. (2023) reference has been added to the bibliography. The screening script and candidate-date file are also deposited in the Zenodo record (R3b, R4).

Technical corrections

T1: Fig. 1 caption — PINCii “squares” vs. actual circles

Reviewer comment

Fig. 1 caption: The sentence about PINCii squares remains in your track-changes manuscript, even though there are no squares in the Figure.

Response

Corrected. PINCii is plotted with circles, differentiated by colour (not squares); the caption now reflects what is actually drawn.

Change in manuscript

Fig. 1 caption: changed “squares” to “circles, differentiated by colour”, so the caption now reads “*PINCii (circles, differentiated by colour)*”.

T2: HYSPLIT and FLEXPART — spell-out and citation

Reviewer comment

HYSPLIT, FLEXPART: these new abbreviations are missing a spelling out and citation.

Response

Both acronyms are now spelled out at first use and accompanied by their primary references. Sect. 3.2 now reads “*HYbrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT; Stein et al., 2015) and the FLEXible PARTicle dispersion model (FLEXPART; Pisso et al., 2019)*”. The corresponding bibliography entries (Stein et al., 2015; Pisso et al., 2019) have been added.

Change in manuscript

Sect. 3.2: HYSPLIT and FLEXPART each spelled out with citation at first use. Two new references (Stein et al., 2015 for HYSPLIT; Pisso et al., 2019 for FLEXPART) added to the bibliography.

T3: Fig. 2 caption — “Savage et al. (2017)” citation form

Reviewer comment

Fig. 2 caption, citation formatting: “are given in (Savage et al., 2017)” should read “are given in Savage et al. (2017)”.

Response

Corrected. The citation form was changed so the rendered caption now reads “Savage et al. (2017)” rather than “(Savage et al., 2017)”. The visual rendering change in the typeset PDF is unambiguous.

Change in manuscript

Fig. 2 caption: citation form changed so the rendered text now reads “Savage et al. (2017)” rather than “(Savage et al., 2017)”.

Closing

We thank Dr. Ladino and Reviewer #2 for the constructive round-2 comments. The corresponding edits are visible in the tracked-changes manuscript; the analysis scripts, multi-seed importance table, dust-screening machinery, and `sessionInfo()` capture are archived at <https://doi.org/10.5281/zenodo.20367476>. We look forward to the topic editor’s reply.