Response to referee 1 (https://doi.org/10.5194/amt-2022-67-RC1)

We thank the reviewer for the comprehensive comments on our manuscript. Below we give our detailed response to each of the comments (shown in italics) and the changes or additions made to the manuscript based on them.

However the manuscript needs some substantial rewriting. The general plan is not presented clearly enough.

After the changes detailed in the responses below the general plan should be clearer.

The abstract states that the processing is under the control of the data owners with a focus on the station level but later it is mentioned that data postprocessing can be done outside the station, and in the conclusion the authors mention multiple stations supervised by the same people. Data harmonization is emphasized which is a very good idea, but is that case the data owner can not really do whatever she/he wants anymore. Data harmonization is a key concept in networks like ICOS where the data processing is centralized leading to the harmonization. The SMEAR concept endorses European infrastructures like ICOS but the authors claim their concept to be different ; however in the conclusion they suggest networking stations together, cross-referencing their data and sharing storage between the stations.

While it is true that higher level networks require you to follow their protocols, many stations/campaigns have measurements that fall outside their scope. In these cases, sharing resources/protocols between such non-infrastructure stations is possible, which is why it was mentioned. SMEARcore is not a replacement for ICOS/ACTRIS protocols and we don't mean to imply it is.

We added clarifying sentences to abstract:

"Secondly, by providing tools for making data interoperable in general instead of harmonizing a particular set of instruments"

"As such it is not meant as a replacement for these infrastructures, but to bring structured data curation to more measurements not covered by them."

We added to Introduction:

", in this paper meaning any process from raw data to products such as end-user data or diagnostics," *to clarify what we mean by analysis. Removed the words "ad hoc" to remove possible confusion.*

"Developing documented workflows for situations not covered by large-scale network protocols is a problem many stations need to solve."

"The interfaces also enable building small networks on top of SMEARcore directly."

"SMEARcore provides flexible and scalable framework that can be applied at instrument, station or multiple station level."

Deleted the following sentence fragment from Workflow:

“and what larger scale infrastructures the station belongs to”

We also added clarifying sentences to conclusions:

“This makes it useful for measurements not controlled by the centralized solutions.”

“This means it is possible to establish smaller networks more easily with the software.”

For example, many measurement campaigns and SMEAR stations have lots of similar measurements which could easily be processed by same codes, cross-referenced if there are problems and stored in the same location, even if they are not part of any large-scale network. Since the SMEARcore could be ran in an internet accessible location, it would even be possible to define a “virtual station” that consists of instruments that are nowhere near each other.

Another point which needs clarification is the use of SMEARCore in the frame of campaigns as it needs some hardware resources.

There is chapter 3.2 about hardware. We added sentences to clarify the sensible minimum requirements (a reasonably modern computer, enough disk space and wired connections between it and measurement computers.):

“In practice this usually means setting up routers and wired ethernet connection between the computers.”

“In general any computer that supports container virtualization and has enough storage can work as the server”

Specific comments:

L18:why a faster installation of new measurement will allow a station to benefit from the experience of SMEARCore ?

We changed faster to structured. Speed is an effect not a cause.

L 33-39: the paragraph focuses on important amount of data and big data but this is not related to the accuracy of mass spectrometry. The raise of number of stations implies indeed more data but not at a single station and SMEARCore focuses on the station level.

We added the sentence: “Doing as much processing as possible at the station can aid in the management of the volume of data.”

L 55: big infrastructures are by default not thought to be interconnected with each other, it is a plus when they do. Presenting the lack of coordination between them as a default is not correct. It is true that processing data can be labor intensive and lack documentation, mentioning this to highlight SMEARCore features is fair but saying that the large infrastructures do not automatize, trace and document their process is incorrect.

We removed the sentence saying that they are not connected. Added sentence: “Developing documented workflows for situations not covered by large-scale network protocols is a problem many stations need to solve.” *The aim, as said, is not to replace any big infrastructures, but to aid in general measurements not covered by them.*

L 101: point two is more a plus than a default requirement.

That ability is pretty much central for much of the existence of SMEARcore. Without analysis capabilities the station cannot effectively monitor itself or the instruments. Now, it is debatable as what counts as "analysis", here we have gone with "can do things with the files it collects".

L 121: indeed but there is no conclusion related to SMEARCore features.

We clarified the implication by adding to end of sentence ", meaning that most instruments can be handled by SMEARcore almost identically."

L 123: « in a conceptual framework » would not be better said as « conceptually »?

Yes, we changed this.

L 138: the required procedure corresponds to the workflow.

We corrected this.

L 139: one workflow or branching workflows?

EC would be a similar parallel, independent workflow. There is no branching here. We added the word "independent" to clarify.

L 150: figure 1 legend, are each box a workflow ? If yes it would be to mention it.

No, they are not workflows in themselves. Each one is a processing step in the workflow. The coloured division boxes are architectural units. We added the sentence to the caption: "The black bordered boxes are steps in the workflow."

L 194: column form time series data is not clear.

Any data of the format (timestamp, datapoint) is columnar form when multiple datapoints are present. We removed the confusing term column form.

L 200: maybe « raw data » will be more clear than « data itself »

We changed this.

L 204: « This way it is ... » is not clear, something like « Multiple views allow to get » may be easier to understand.

We changed this.

L 219: « restrict instruments so they can access only their own data.» is not clear.

We extended and reformulated the sentence to: "They also allow us to isolate instruments so that each has their own folders, and they cannot even accidentally overwrite other data."

So, instrument A can write their files even if instrument B has the same name of raw files.

L 222:Apache Airflow is only use for the analysis workflows, what about the others?

Analysis is the only part with complex workflows. The data collection operates on a simple loop. Nothing in principle stops one from using Airflow for other parts, which would be a possible future development.

L 252-253: « SMEARest » SMEAR Estonia? Grafana processes the metadata? The whole sentence is not really clear.

We split the sentence into two parts. Grafana offers visualizations (as normal), as well as some additional metadata about the collection process itself. It does no processing by default (it's possible to define functions, but that is beside the point).

L 262-274: too long for the purpose of the manuscript.

We removed unnecessary details. The removal of figure 6 and supporting sentences also shortens this section. The purpose is to show how having access to the data of multiple instruments helps in interpreting the situation.

Figure 3 legend: the information of the custom plugin is interesting and it would be more appropriate to move it to the core of the manuscript.

We moved it to the chapter discussing the plot.

Figure 5 and 6 are redundant, one would be enough with the legend specifying that inorganic data can be presented the same way.

We removed previous Figure 6 (adjusted other figure numbers and fixed figure labels in text) and added a note about inorganic data in the caption.

L 306: remove « but », an email is also sent ot the operator.

We replaced ",but also" with "and".

L 318: remark on SMEARCore, it should trace the removal of the instruments from the station in order not to display false status.

It's impossible to resolve a missing instrument without operator given metadata. In this case one can also just remove the instruments from the workflows/views. So, this is not really something SMEARcore can trace automatically.

Figure 8: may be it can indicate how far in the pas the data are available.

That is a possible improvement that was not implemented during this study.

L 338: « parallel implementation » is not very clear.

The station operated normally at the same time; we did not replace any functions. Instead, we had our own analysis pipeline. We removed the word parallel, since this is not a necessary detail for the manuscript.

L 346: « running the workflows as graphs in Airflow » sentence is a bit to technical. Maybe something like « the workflows in Airflow are defined by graphs ».

We changed it to: "The analysis workflows in Airflow are defined by graphs"