https://doi.org/10.5194/egusphere-2024-1469-AC2

Reply To The Editor

Author Response:
We would like to thank the reviewers and editor for taking the necessary time to review the manuscript. We are sincerely grateful for all their valuable comments, which helped us to improve the quality of the manuscript.

RC1 Comments, responses and changes

Reviewer Comment:
This manuscript addresses the very important subject of data management for large environmental datasets. The items discussed in this manuscript are often overlooked which can lead to confusion during publication and future data analysis. The authors have obviously put a lot of thought and experience into the design of the data management platform.

There are three overarching items that should be addressed before publication. The first is that acronyms need to be defined upon first use and then their use should be consistent throughout the manuscript. The second is that some of the sentences are rather long and difficult to follow. Technical editing for clarity is recommended. The third is that some of the figures seem redundant or overly complex and simplification would greatly improve the readability of the manuscript. Some specific comments are provided in the attached pdf.

Author Response:
We thank the reviewer for the detailed assessment and valuable comments. Below we address the items from RC1 and the annotated supplement provided by the reviewer.

We agree with the reviewer that the three overarching items must be addressed.

Author Changes:
We (1) revised the text to improve the use of acronyms, (2) made technical editing on long sentences and (3) improved the captions and presentation of figures. Details are provided below.

Responses to the comments in the PDF supplement to RC1:

Lines 58-64:
"This makes assessments time-sensitive, diagnosing network operational status critical, to inform models, to make decisions on resource-intensive field deployments (e.g. radiosoundings, tracer releases), and to make dynamic adjustments to network design (…;…)."

Reviewer Comment:
This sentence does not make sense.

Author Response:
We agree with the reviewer and changed the sentence.

Author Changes:
"The dynamics make data assessments time-sensitive and network operational status diagnostics a critical task, as intermediate results become indispensable for informing models, making decisions on resource-intensive field deployments (e.g. radiosoundings, tracer releases) and making dynamic adjustments to network design (...;...)"

Line 86:
"The different objectives in Module A to C require targeted and specific observational strategies, but all require consistent data management, documentation and quality control processes. (Module D)."

Reviewer Comment:
Strikethrough text.

Author Response:
We agree with the reviewer.

Author Changes:
"The different objectives in Module A to C require targeted and specific observational strategies, but all require consistent data management, documentation and quality control processes (Module D)."

Line 94:
"Hence, there are fixed deployments and mobile measurements, sensors with multiple uses, need for near real-time data e.g., during intensive observation periods (IOPs), changing configurations, deployments of varying duration (e.g., hours, days, months, years)."

Reviewer Comment:
put inside parentheses, insertion "and"

Author Response:
We agree with the reviewer.

Author Changes:
"*Hence, there are fixed deployments and mobile measurements, sensors with multiple uses, need for near real-time data (e.g., during intensive observation periods (IOPs)), and changing configurations, deployments of varying duration (e.g., hours, days, months, years). This is managed by multiple people with different responsibilities, roles and backgrounds."*

Line 97:
"The system currently ingests the order of 10^9 datapoints per day from about 100+ different stations and approximately 1000 sensors in five different cities."

Reviewer Comment:
on the order of

Author Response:
We agree with the reviewer.

Author Changes:
"*The system currently ingests on the order of data points per day from about 100+ different stations and approximately 1000 sensors in five different cities.*"

Line 112:
"The physical, logical, and organisational context of the field observations
are defined early in the data management process (Table 1) because of the latter, from planning to collection to publication."

Reviewer Comment:
This does not make sense

Author Response:
We agree with the reviewer.

Author Changes:
"Because of the latter, the physical, logical, and organisational context of the field observations are defined early in the data management process (Table 1), and amended with updates from planning to collection to publication."

Lines 117-120:
"Examples of the operational relational queries, using a graphical user interface (GUI), include finding: if an identical instrument exists in storage or another deployment; all instruments at a location; all mobile-phone SIM cards linked to a data plan; all calibration sheets for a given sensor, and warranty documents to facilitate sending an instrument back to a manufacturer for service."

Reviewer Comment:
I would suggest putting this information into a separate sentence. These examples may fit better into the explanation if they are numbered.

Author Response:
We agree with the reviewer.

Author Changes:
"*The operational relational queries are supported by graphical user interfaces (GUI) with shortcuts for specific summaries. Those help find, e.g., (1) if an identical instrument model exists in storage or in another deployment, (2) all instruments at a location, (3) all mobile-phone SIM cards linked to a data plan (4) all calibration sheets and warranty documents for a given sensor to facilitate sending an instrument back to a manufacturer for service.*"

Lines 129-131:
"The hierarchy of spatial information for a "station" (or "site") is from geographic coordinates, to a local Cartesian or polar coordinate system (Figure A3) to explain details in a complex setting such as a roof (Figure A1) or within a building (Figure A2)."

Reviewer Comment:
I would suggest breaking this into 2-3 sentences to make the meaning more clear.

Author Response:
We agree with the reviewer.

Author Changes:
"The hierarchy of spatial information for a "station" (or "site") starts with geographic coordinates of a point on a representative surface. These, which can be determined accurately in advance to assure suitability for sampling by (airborne) remote sensing. The "system" and "sensor" are measured in relation to the "site" in a local Cartesian or polar coordinate system (Figure A3). This explains details in a complex setting such as a roof (Figure A1), a street canyon or within a building (Figure A2)."

Line 131:
"The database relationships allow identical sensor replacement (e.g., AWS), without needing to modify any of this spatial information."

Reviewer Comment:
Amazon Web Services? Please make sure to identify all acronyms upon first usage.

Author Response:
"AWS" referred to an "Automatic Weather Station". We agree with the reviewer. In this line the acronym adds confusion. The proposed change is to remove the "(e.g., AWS)" part. The three-character acronyms of system groups are explained elsewhere in the text.

Author Changes:
"The database relationships allow identical sensor replacement, without needing to modify any of this spatial information."

Line 148:
2.2 Convention

Reviewer Comment:
I would also recommend that the authors comment on time convention. The best practice, in my opinion, is to have two columns for time, one with the time beginning of the measurement period and one with the ending time of th measurement period.

Author Response:
We thank the reviewer for the suggestion.

An accurate time convention is critical. The cf-conventions-1.10 offers comprehensive and clearly defined options to describe and encode the start and end of intervals of time (or any other dimensions), and declare relevant attributes of the 'cell' value. For many of the data products, the data are 'point' samples stored in their original sampling resolution. This is specified as an attribute to the variable (cell_method: "time: point"). Upon any aggregation, along the time dimension or any spatial dimension, the bounds can be (must be) declared and the cell_method attribute for those variables updated accordingly. The convention further reserves suffixes for aggregation to be added to the variable name, such as _mean, _maximum and so on.

Author Changes:
No change to the text.

Lines 149-152:
"Building on existing data conventions and standards can enhance data usage. In urbisphere, we use the climate and forecast (CF) metadata conventions (cf-conventions-1.10, or CF hereafter; see Hassell et al., 2017) API for NetCDF, with extensions often used in the urban research community (Scherer et al., 2019) and consistent with prior campaigns, model applications, third-party software tools, common with the campaign's instruments and project-specific production needs."

Reviewer Comment:
This is a long sentence, I suggest breaking it into two.

Author Response:
We agree with the reviewer.

Author Changes:
"In urbisphere, we use the climate and forecast (CF) metadata conventions (cf-conventions-1.10, or CF hereafter; see Hassell et al., 2017) API for NetCDF, with extensions
often used in the urban research community (Scherer et al., 2019). The use is consistent with prior campaigns, model applications, third-party software tools, and common with the campaign's instruments and project-specific production needs."

Line 157:
"Many instruments and data recorders encode sensor data records to proprietary formats, which based on user options may provide calibrated and aggregated data records with sensor diagnostics."

Reviewer Comment:
This has not been my experience with a variety of commercial instruments. In all cases, the data can be exported as a csv. If this has been the case for many instruments used in the urbisphere study, perhaps this study can also comment on how to save data to csv files to facilitate data accessibility.

Author Response:
Thank you for sharing experience and insights.

The usefulness of CSV is limited to point measurements (Table 3) and some aggregated data products. Typical point measurement instruments, such as an automatic weather station, output CSV as a default option. For most other instrument groups it is not (or no longer) meaningful to choose CSV, particularly when diagnostic data and summary computations need to be saved alongside the sampled measurements. The proprietary formats, sometimes with complex structures in delimited text files, sometimes hierarchical NetCDF, JSON or XML, must be ingested before they can be exported to CSV. We assist researchers with the ingestion process by preparing L0 datasets, which are identical in content to the RAW source, but structured in a robust and consistent manner across all feature types of the measurement (point, path, area).

CSV is still a first-class supported output of the current generation modules that are capable of interacting with NetCDF, the conversion and export to CSV is often as simple as a one-line command. We do not make any APIs for CSV (or GeoJSON) available to partners as the data structure is too simple and flat to maintain the essential metadata (attribution, terms of use, history, references). Zarr, NetCDF and JSON are therefore preferred formats.

Author Changes:
No change to the text.

Line 167:
"The shared vocabulary facilitates efficient queries and multiple benefits."

Reviewer Comment:
What are the multiple benefits?

Author Response:
We agree with the reviewer. It is currently not clear that the next sentences list examples of some of the benefits.

Author Changes:
Replace "and multiple benefits" with "and benefits for the machine-operability of the data."

Line 170:
"Existing community software tools work with NetCDF and CF-related vocabulary definitions, including modules to perform programmatic conversion of units of quantities (e.g., the UDUNITS module; Hassell et al., 2017)."

Reviewer Comment:
Is this referring to concentrations?

Author Response:
We agree with the reviewer. We thank the reviewer for pointing to this unclear part of the text.

Author Changes:
"Existing community software tools work with NetCDF and CF-related vocabulary definitions, including modules to perform programmatic conversion of units (e.g., the UDUNITS module; Hassell et al., 2017)."

Lines 174-175:
"Measurement of, e.g., near-surface air temperature in an urban area may require different siting requirements than standard WMO regional scale weather, with neither satisfying the other (Stewart, 2011; WMO, 2023)."

Reviewer Comment:
This sentence is awkward. Suggest rephrasing to something like: "For example, the measurement of near-surface air temperature in an urban area may require different siting requirements than standard WMO regional scale weather measurements."

Author Response:
We agree with the reviewer.

Author Changes:
"For example, the measurement of near-surface air temperature in an urban area may require different siting requirements than standard WMO regional scale weather measurements (Stewart, 2011; WMO, 2023)."

Line 177:
"Most deployments are at fixed locations, but as surroundings change, regular review of deployment configurations are required, and therefore metadata adjustments."

Reviewer Comment:
'and therefore metadata adjustments' what? This either needs to be moved to before 'are required' or additional information is needed.

Author Response:
We agree with the reviewer.

Author Changes:
"Most deployments are at fixed locations, but as surroundings change, regular review of deployment configurations are required, and time-specific amendments are needed to the metadata."

Line 177:
"Some instruments have accurate clocks and sensors to self-determine location and orientation, providing metadata details (e.g., radio-sounding sensors)."

Reviewer Comment:
A clock will not determine location and orientation. Please identify what sensors are often used for these purposes.

Author Response:
We agree with the reviewer.

Author Changes:
"Some instruments have accurate clocks and sensors to self-determine location and orientation, providing metadata as a separate time-series in the data (e.g., pressure and GPS sensors on radio-sounding systems; motor-drive position and inclinometer on lidar systems)."

Line 178:
"Other deployments may require frequent offsets to be determined or determined continuously relative to a (local) reference."

Reviewer Comment:
offsets for what?

Author Response:
We agree with the reviewer. This is not clearly formulated.

Author Changes:
"Other deployments may require the sensor viewpoint and orientation as well as time offsets to be measured regularly or determined continuously relative to a (local) reference."

Line 188:
"All systems have GPS time or internet reference time services, and all data during urbisphere campaigns are generally recorded in UTC (Appendix A1)."

Reviewer Comment:
Were there any exceptions to the recording in UTC?

Author Response:
We agree with the reviewer. This was not specified properly.

UTC, or more precisely for the European campaigns, GMT or UTC+0000, reports no daylight saving offset that can cause confusion in recordings of timestamps in a format that is unaware of time zone. All computers and clocks were set to use UTC or GMT. Hence, any timezone-unaware record in the data would imply UTC. This was validated meticulously for each deployed system by the field teams. Time validation was further supported by an online index of latest files arriving to the remote server, highlighting the modification times and filename timestamps (Figure C2).

Author Changes:
All systems have GPS time or internet reference time services, and all data during urbisphere campaigns are recorded in UTC or the GMT/UTC+0000 locale without daylight saving for systems with time-zone unaware recording of timestamps in line with common practice in meteorology (Appendix A1)."

Line 195:
"Siting for all sensors need to be representative of what people and room are likely to experience, while still allowing the rooms to be used in its intended way."

Reviewer Comment:
Strikethrough text.

Author Response:
We agree with the reviewer.

Author Changes:
The word "rooms" is replaced by "room".
"Siting for all sensors need to be representative of what people and room are likely to experience, while still allowing the room to be used in its intended way."

Line 217:
"Key features of the physical and logical networks (Section 3) are presented to organisational aspects (Section 4)."

Reviewer Comment:
Strikethrough text.

Author Response:
We agree with the reviewer.

Author Changes:
Replace "(Section 3)" with "(this section)"
"Key features of the physical and logical networks (this section) are presented to organisational aspects (Section 4)."

Lines 256-257:
"Data are stored in the instrument-provided formats, which may be custom text records (encoded information in a proprietary format, e.g., TOA5, or a defined schema, e.g., XML), where new records are appended as lines to a file with a header that contains metadata and a column description. "

Reviewer Comment:
TOA5 is very easy to replicate and while it is one of the default header formats for Campbell Scientific loggers, I would not necessarily call it proprietary.

Author Response:
We agree with the reviewer that TOA5 is easy to replicate, but the header structure is determined by Campbell Scientific. To maintain compatibility, the header information, including the number of header lines, sample types, name format and units, is only customisable to a very limited extent. TOA5 is very useful, and documented, but proprietary to the point of being designed to the specifications of Campbell Scientific loggers.

Author Changes:
No change to the text.

Line 266:
"As a research network, temporary and long-term network outages must be accounted for in the design, requiring sufficient local data storage on or near the instrument, as well as methods to resume data transmission after an outage."

Reviewer Comment:
I would also add that data transmission of records that were collected during the temporary network outage is important.

Author Response:
We agree with the reviewer.

This is implied, the data recorded since the start of the network outage need to be transferred, ideally automatically. This requires methods that identify what is missing on the remote storage location, and skip redundant uploading to save bandwidth.

Author Changes:
"As a research network, temporary and long-term network outages must be accounted for in the design, requiring sufficient local data storage on or near the instrument, as well as methods to resume data transmission after an outage. The data recorded since the start of the network outage need to be transferred, ideally automatically. This requires methods that identify what is missing on the remote storage location, and skip redundant uploading to save bandwidth."

Lines 281-282:
"Custom software is used to configure the RSYNC client software and set retention periods for data transmission (Morrison, 2022)."

Reviewer Comment:
What is the implication of using custom software for this in terms of accessibility and the ability to replicate the network described here?

Author Response:
We thank the reviewer for the comment, it is indeed a valid point of critique about the (re-)useability of custom software.

RSYNC is a client-server tool. The use of RSYNC depends on SSH and RSYNC installed on both the client and server. On linux systems both are standard. The operation of RSYNC, SSH and linux systems requires experience. This solution was known and available to us when we designed and implemented the back-end (in a very short time, a few weeks). Operationally, it required the use of so-called chroot-jailed SSH accounts, in which rsync is one of only a few functionalities available.

More advanced, configurable and replicable options for data transmission are available but we cannot report experience with those. We briefly considered SyncThing and Rclone as potential open source alternatives and those looked promising. Meanwhile more scalable options are being made available to us, such as SFTPgo tied to cloud storage services (e.g, Amazon Simple Storage Service/S3), but we have still to gather experience with those platforms, particularly when working with the often outdated and legacy connectivity of data loggers. Such services, i.e., managed by central IT at the host, do promise to reduce the overhead in "data engineering" for projects. Alternative to file transfers we envisioned a REST API, which we could not develop due to time constraints.

The custom software (Morrison, 2022) is open source and developed to simplify and replicate the configuration of RSYNC as well as any scheduled execution of RSYNC and housekeeping tasks on the deployed local network node (the data logger). Limited accessibility was the main driver to develop this helper tool, as current mobile phone networks are blocking direct access to the local nodes from any remote node on the internet. Although working with a VPN network solves this limitation, the design to push data to a remote node, rather than to pull data from a remote network node, is simple and proved adequate for most use cases. Remote access software was used for troubleshooting.

However, more important is that a local storage and network node must operate with or without online access. Housekeeping tasks must be optimized to work without remote access or management, storing and safeguarding data (for as long as local data are unique copies). Furthermore, field staff must be able to collect data manually, at all times (i.e., local network access, typically without VPN).

Author Changes:
No change to the text.

Lines 290-292:
"VPN is available through some routers (e.g., model RUT240 and Teltonika services, Teltonika Network, Kaunas, Lithuania), or commercial remote desktop software (e.g., AnyDesk Software GmbH, Stuttgart, Germany)."

Reviewer Comment:
It is also possible to set up a virtual machine in AWS that is within a network.

Author Response:
We thank the reviewer for the suggestion. Unfortunately, we cannot report our own experience with the use or virtual machines in Amazon Web Services within a network.

Author Changes:
No change to the text.

Line 382:
"The implementation primarily uses: freely-available software tools, established services for storing research data, and community adopted conventions."

Reviewer Comment:
It would be of interest to other campaign managers and PIs as to the cost associated with the tools.

Author Response:
We agree with the reviewer and propose to add details about the cost.

Author Changes:
The following lines were added to the "3.7 Operational Costs" section.
"The virtual hardware is provided by the host (approx. EUR 500-1000 per year) with a one-time purchase of data storage units (approx. EUR 40.000). The software tools are open source, except for remote access software license (approx. EUR 250 per year). All tools need to be configured, for which the personnel costs include a data scientist, as well as a researcher, a field technician and research assistants for each campaign."

Table 3 / Lines 396-398:
"Many instruments have dynamic source areas either because of meteorological conditions or if the sensor is mounted on a mobile platform"

Comments:
Are these examples from a specific campaign? Can additional classifications be added?

Author Response:
These examples are from the current and past campaigns in the described project. The motivation here is not to be restrictive, but to reserve acronyms for 2- 3- 4- and 6-character uppercase acronyms to, city, system group, station identifier and combined city-station identifier reference, respectively. We found the consistent use of those formats in metadata, communication and publication helpful.

Author Changes:
Text added to Appendix B1:
"The motivation for the use of definition rules for acronyms is not to be restrictive, but to reserve acronyms for 2- 3- 4- and 6-character uppercase acronyms to, city, system group (Table 3), station identifier and combined city-station identifier reference, respectively. We found the consistent use of those formats in metadata, communication and publication helpful."

Figure 7 / Lines 421-424:
"VDI 3786 (VDI, 2013) quality control indicators applied to the Freiburg AWS network (rows) during a storm (11 to 12 Jul 2023) for all variables (precipitation pr_rate; air temperature ta; relative humidity hur; wind speed ws; wind direction wd; station pressure plev; incoming short wave (or global) radiation rds) with quality flags shown: long = "bad", short = missing."

Reviewer Comment:
Is this figure really necessary in the main text? IT is difficult to discern all the items presented here. I would recommend that the authors reconsider this figure's presentation."

Author Response:
We thank the reviewer for the critical note and agree that the presentation of the figure would either need to be improved substantially, or the figure be removed.

Author Changes:
Both Figure 6 and 7 were moved from the main text to Appendix B, subsection Quality Control, and accompanied with additional explanatory to improve the presentation.

"An automated assessment of typical meteorological variables was implemented using the Verein Deutscher Ingenieure (VDI) guidelines for meteorological observations (VDI, 2013). The guidelines provide threshold values for the range, rate of change (absolute deviation) and duration of steady-state (stationarity duration) for a number of variables. The threshold values are specified for different averaging times. Testing the data with these threshold values determines for each data point if the quality is good, ambiguous or poor. The procedure for change rate and steady-state calculations require data before and after a data point to be available and the computations involve repeated averaging at different time intervals, which incurs computational costs and complexity. Additional care is needed to assure the units between the data and threshold values match.

The output of the quality control is a new dataset with the same time dimension as the original data. The result can be summarized into an ensemble quality control indicator for specific variables, which can be useful for evaluation and masking data points before further use (Figure B1). By combining all quality control output for a network of sensors and multiple variables, outliers and trends can be assessed (Figure B2). The example describes a passing storm on the evening of 11 Jul 2023, registering (1) a rapid humidity change at most locations, (2) high variability in wind speed and/or wind direction at some locations depending on the orientation of the street canyon to the wind (3) a possible malfunction in the precipitation sensor at FRLAND (4) a possible time offset of the system at FRCHEM, and (5) an unspecified technical issue that affects the data delivery at FRTECH. The information should be considered indicative, as it can reveal both natural changes and technical problems, but can be further supported by spatial statistics (not implemented here) and field reports.

Figure 12 / Lines 436-437:
"Pathways to making data publications public accessible require the different data governance issues to be formalized early, allowing access to be rapid after any restrictions."

Reviewer Comment:
Could this caption please be updated to include information about what the figure is showing? And how is this different than previous figures showing the data pathway?

Author Response:
We agree with the reviewer that the caption is insufficiently describing the figure. The figure describes the data governance aspects.

Author Changes:
"Figure 12. Pathways to making data publicly accessible require different data governance policies (yellow boxes) to be formalized early, allowing public access to be rapid and restrictions to be mitigated. Open access of collected data (RAW) is an option without such policies (bottom pathway), and embargoed release can be agreed with policies in place (center pathway), but the curation and analysis work (L0-L2) can involve intellectual ownership and personal interests that may otherwise lead to delays in open data publication (top pathway)."

Figure 13 / Lines 438-439:
"An info-graphics using a switchboard analogy is used to communicate where data is uploaded, where data can be downloaded and where data streams are being managed or modified."

Reviewer Comment:
While a beautiful figure, does this add value to the manuscript? What are the different colors of lines? How is this different than the previous figures?

Author Response:
We thank the reviewer for the comment

Currently, the data infrastructure is used by four independent projects, and campaigns within those projects. Each with their own requirements and user access management. Figure 13 was used to show newcomers the five main terms used in the documentation (issue tracker, email, otherwise) with details about the access to their 'archive'.

Figure 8 shows the data infrastructure components and data flow, Figure 12 shows the data governance pathways. Figure 13 shows both, but in an oversimplified manner. As organizational and logical networks need to be meshed into one, the data system is more complex; depending on the role, depending on the task, different access points exist to the data, both on the input side as on the output side.

The annotation is indeed missing. Intended to show were RAW in pink, L0-L2 in blue, in accordance to the other figures in the manuscript.

Author Changes:
Annotations were added to the colored lines.

RC2 Comments, responses and changes

General comments:

This paper describes at high- to mid-levels a framework for data-centric management of extremely heterogeneous observational sensor networks based in urban environments. The authors describe the motivating scientific drivers as requiring both fixed/long-term as well as mobile/short-duration campaigns. This set of requirements is complex and ambitious, representing the cutting edge challenges of applying the Internet of Things to observational science. This paper makes important contributions to approaching these challenges with a systematic and modular architecture, with attention to the scientific necessities of data integrity, traceable quality control, and data interoperability. The paper narrative is well-organized and well-written, with no real issues in terms of language, communication, or structure. The figures and tables are generally clear and communicative of key concepts. Some of the related low-level technical implementations and solutions remain vague, as well as the scale/implementation of the information technology infrastructure and related engineering roles. The paper is certainly valuable and makes a community contribution as-is (with excellent descriptions of sensor metadata and data workflow frameworks), however the authors should consider a section (or a separate paper) detailing the back-end infrastructure solution, the associated engineering team, and any implemented development-operations steady-state workflows that support long-term resilience and consistency. Major observational networks in the publicly-funded science sphere have always struggled with prioritizing the back-end engineering to avoid "technical debt" on the decadal scale, and multi-generational operation of observational science now depends on the development-operations architecture and associated engineering management.

Author Response:
We thank the reviewer for taking time to review the work, for providing comments and for sharing relevant insights on the scalability and organization of complex data-centric research. We address major and minor comments in detail below. Some discussion points are discussed in a separate response.

The work is centered around a soft-funding staffed academic research project with an operational window of a few years, not decadal, hence, the "technical debt" is almost entirely guaranteed by design. None of the current participants, beside the PIs, should be expected to be involved to continue development of the presented data-centric solutions after 2027. The long-term resilience lies in the generation of long-term persistent, consistent sets of data and metadata, i.e., on Zenodo.

We hope our report is of help to the urban environment community and those groups facing similar challenges for short-term ad-hoc deployments. We agree with the reviewer that the low-level back-end engineering solutions are of critical importance to the reuse and we aim to share those in more detail (in a separate paper, Github, or technical report).

Reviewer Comment:
Specific comments: Note: many of the following comments are made with respect to the potential of this manuscript to contribute to improving the observational community's concept of a model and modern large-scale heterogeneous framework, which the authors state is their objective.

Line 25:
Key Figure or Graphical Abstract

Reviewer Comment:
It may be an issue of nuance or language use, but "data management" falls short of representing the foundational technology platform/architecture of end-to-end data acquisition-transport-staging-processing-production-dissemination-archival. Modern technology-driven business applications recognize clear differences between "data science" (science requirements, analysis, products), "data management" (logical requirements, policies, workflow design, quality control), and "data engineering" (software & hardware architecture, software development-operations, performance management, end-to-end user/security/network design/implementation). Longevity, repeatability, and resiliency of a complex heterogeneous digital system requires teams of expertise focused on each of these separate roles, and (all but the largest) scientific operations

struggle with recognizing and provisioning them. This key figure/graphical abstract over-simplifies the technical foundations of modern and future observational science at scale.

Author Response:
We thank the reviewer for the comment.

This is indeed an issue of nuance. The graphical abstract informs on a development we identified: towards larger and more heterogeneous networks of observations that need to be integrated. As the complexity increases, the data stream must be managed in different ways.

A small team is tasked to 'manage' the data, combining engineering, management and science practice. Keeping the roles separate would be helpful, but in reality the provisioning for the scientific operations is limited to very few people, therefore the lines between the roles blur.

We agree with the reviewer and propose to include a description of the roles "data engineering", "data management" and "data science", separately and in detail.

Author Changes:
No change to the text (in line 25).

Change to the text (Section 4):

"We should further recognise the responsibilities for (1) data science (i.e., scientific requirements, analysis, products), (2) data management (i.e., logical requirements, policies, workflow design, quality control) and (3) data infrastructure engineering (i.e., software and hardware architecture, software development, operations, performance management, end-to-end user/security/network implementation). There are clear differences between these responsibilities, and having experts focus on each separately can arguably improve data system resilience and longevity. However, setting up teams of data experts is not common in soft-funded academic projects engaged in short-term collaborative observational campaigns, and responsibilities end up being carried by few people (see Section 3.7)."

Change to the acknowledgements:

Listing the reviewer in the Acknowledgements by name for providing valuable contribution.

"We acknowledge an anonymous reviewer and Scotty Strachan for their important contribution to the quality of this publication."

Line 97:
"The system currently ingests on the order of 10^9 datapoints per day from about 100+ different stations and approximately 1000 sensors in five different cities."

Reviewer Comment:
An observational network of this scale requires significant hands-on quality review and control in order to ensure production of scientifically-valid data. An estimate or description (somewhere in the manuscript) of the number of expert/hours/day that are needed to perform continuous data review to maintain consistent quality would be valuable.

Author Response:
The project's primary objective was the collection of the data, and metadata. The quality review and curation is supported by automatic quality assessments and machine-operabale data structures (shared vocabulary, units), and linked to databases of Events (field log, maintenance logs, review log, etc). Further scientific scrutiny is required for validation and depends on the instrument group and application. At this point we cannot estimate how much time the data users spent on (continued) data review tasks.

Author Changes:
No change to the text.

Lines 110-147:
Section 2.1 Metadata

Reviewer Comment:
The authors are to be commended on their attention to detail and variety in the metadata design. It would also be good for the community to know more detail about how these properties get populated (for example, use-case workflows for different categories of sensors, both in the setup/deployment phase as well as the maintenance phase). Useful details would be the method of metadata access/modification, the expected time requirements of the operators, etc. The entire manuscript describes the intended design, but reveals very little in terms of day-to-day function.

Author Response:
The use cases are centered around the individual campaigns, for which the workflows and functions are published separately (i.e., as supplements to research papers, and as records in the Zenodo community urbisphere).

Author Changes:
No change to the text.

Lines 110-147; 199-209:
Section 2.1 Metadata and Section 2.3 Operational management.

Reviewer Comment:
Without knowing much about how the various databases are interacted with by operations and applications, I'll mention that there are other architectures and models for the overall data/metadata management that are emerging in the community. For example, a cloud-hosted modular multi-network sensor data system is in operation and development for several U.S. academic monitoring networks (https://dendra.science/; https://essopenarchive.org/doi/full/10.1002/essoar.10501341.1), that utilizes metadata and QC levels as "annotations" that are applied on data extraction from the RAW/L0 DB. This is a very different approach from the classic "level-version-copy" method generally used by the community, but has potential for better long-term data efficiency/scalability, semi-automated QC interface/entry improvements, and scientific stability.

Author Response:
We thank the reviewer for the comment.

We do currently serve other projects, which are also depending on using legacy solutions instead of the more modern architectures of choice that could not be configured (in time) to handle certain data structures, volumes or protocols.

It is highly valuable to see community driven solutions for the operational management of networks of instruments. In the described project we depend on some of those initiatives, such as AERONET, PANAME and the PhenoCam Network, which have specialized on particular instrument groups. The heterogeneity of the instrumentation did weigh in on the choice to work with an architecture that requires less integration of metadata and data. A choice was made to use cf-conventions-1.10 and built upon the outcomes of the Urban Climate Under Change data standard [1], to allow both observation and model data to use features of NetCDF.

In our experience there are vocabulary translation challenges in any ingestion system, and then again towards the model application. It will take time to understand, configure and (maybe) develop applications for new instrument groups prior to the start of a campaign.

We address the annotation of data in a Final Response, for completion:

We applaud a community shift towards the annotation of data and recommend working accordingly. In our case study, only the near-time (L2R) production level products have automated quality control applied as a filter to the data, as end-users typically do not have the opportunity or context to evaluate the quality otherwise before use (e.g., when querying only the latest samples). However, the source data (RAW, L1 and/or L2) are made available in full resolution, with a separate quality control data set in coordinates aligned with the data.

Derivative data products generally apply quality filtering, imputation models and aggregation statistics or computations, and are typically published separately, i.e., as supplement to research papers that describe the modifications in detail.

Lines 219-229:
"The data infrastructure combines the interest of data safe-keeping with data accessibility (Figure 8). Keeping data secure is a primary project deliverable and involves basic protection against unknown malicious actors and protection against accidental data loss. Within the data architecture, a central operational archive is maintained on a suitably large storage volume (larger than 50 Tb logical volumes on RAID storage units). A replica of the data is maintained on an identical storage unit in a different building (geo-redundant backup), with additional daily backups on enterprise storage services (on and off campus). The data infrastructure uses virtualised computing hardware. Virtualization makes it possible to isolate critical functions without the need to expand physical hardware, and allows the data infrastructure to scale dynamically as needed. The critical functions include a remote access node for all uploads from local field stations ("gateway"), a remote access node for metadata databases and related web-interfaces ("status"), a remote storage node for archival of data and public access nodes ("workstations") for monitoring, computations and user access to data (Figure 8)."

Reviewer Comment:
As previously mentioned, the back-end infrastructure solution is of key interest to the long-term observatory community. Information technology architecture, engineering design, features, and required human/machine resources for steady-state operations would be of significant value. Initiatives such as https://dendra.science/ (alos show us how helpful

Author Response:
Platforms that provide access to data and metadata are valuable, perhaps critical, for operational management of observatories. But we identified that short-term deployments offer additional challenges.

Some software tools, legacy and community-maintained, depend on using original files as input. Our approach for data collection during those short-term deployments is to maintain as much of the instrument-native instrument output as possible. A steady state may not be easy to reach during limited-time campaigns, when configurations are changed to serve different research goals. Saving the original data files as-is makes the collection independent of data processing platforms and may otherwise be required by data safekeeping policies.

For short-term observatories the tools for diagnostics and visualization ideally are steady-state, but often end up to be on-demand in an interactive computing environment (ICE) of choice. A solution here is to work with remote (virtual machine) workstations, where ICE code can be developed, automated, and output shared with others in ad-hoc online Apps and APIs.

The text is not currently highlighting the required low-level solutions, as the reviewer states, and we propose to provide those in a separate publication.

Author Changes:
No change to the text.

Lines 230-252:
Section 3.2 Access

Reviewer Comment:
This sounds complicated to manage and ensure integrity, and it appears that there are many case-based exceptions and manual permissions-related processes to be handled. This will prove difficult to align with emerging cybersecurity requirements, and likely complicates the process of tracking data provenance. In general, friction in any process for researchers to interact with data and technology solutions means less participation and longer time-to-science, so perhaps the architectural team will keep this whole access management process high on its priority list to automate and reduce complexity.

Author Response:
We thank the reviewer for the comment and insights.

As the reviewer identified, the system is complicated to manage, particularly due to permissions related processes and exceptions. Most partner researchers are aware that those measures are not in place to obstruct their data access, but to fulfill cybersecurity requirements. A main objective of the project is to collect and safeguard the observational data. Tracking data provenance is a concern which the project aims to overcome with the timely publication of the original data and metadata.

Author Changes:
No change to the text.

Lines 278-287:
"The logical network uses industry-standard protocols for the transmission of data files (i.e., FTP, SFTP and RSYNC through SSH). SFTP adds a secure authentication and encryption layer to the transfer (cf. FTP), whereas RSYNC adds incremental, compressed and validated data transfer (cf. SFTP). RSYNC is preferred, as it allows reliable recovery of incomplete or failed transfers with limited bandwidth overhead on the logical network. Custom software is used to configure the RSYNC client software and set retention periods for data transmission (Morrison, 2022). The synchronisation of data between storage locations also relies on RSYNC (i.e., as transport method for Lsyncd). We find the FTP protocol is no longer fully supported by all mobile phone network carriers. As some data loggers (e.g., model CR1000X, Campbell Scientific, Logan, Utah, USA) use alternative protocols, the upload server is configured to allow legacy authentication methods for SFTP connection. The flexibility to make such server-side adjustment to the configuration underpins why ad-hoc research data collection benefits from a dedicated, custom data infrastructure (Figure 9)."

Reviewer Comment:
This somewhat legacy approach to transport protocols (compared to modern large-scale business data systems in industry) is understandable but also difficult to scale and manage. Are there any considerations of using a logical network architecture based on store-and-forward publish-subscribe data aggregation protocols?

Author Response:
We thank the reviewer for the comment.

The use of legacy methods was not avoidable. We would have implemented a more modern protocol and addition of IoT-based nodes if the global market would have allowed us to source the required components in sufficient quantities in 2021/2022 (e.g., Raspberry PI). Without those, the adoption of more modern protocols was hindered by technical limitations, e.g, data loggers with outdated SSL authentication, and an understandable hesitation by partners and owners to make software changes to systems that 'just work'.

Author Changes:
No change to the text.

Lines 350-358:
"Examples of different type of roles of users of the system include (Figure 2, Figure 5,Figure 9, Figure 10, Table A2).

– Principal Investigator: executive responsibility for all scientific activities, campaigns, data and peer-reviewed publications and priorities;

– Campaign Manager: lead for all aspects of a particular campaign (city);

– Data Manager: lead for data infrastructure supporting campaign teams;

– Publication Manager: responsible for a data publication process;

– Researcher: responsible for a particular data production line or instrument group;

– Field Operator: responsible for deployment and maintenance."

Reviewer Comment:
I don't see separate roles for the "data management" from the "infrastructure engineering" (see my first comment for Line 25). Industry standard for highly-reliable and long-term digital infrastructure systems place greater emphasis, cost, and scale on the engineering teams as opposed to the scientific and management/policy teams.

Author Response:
There is great emphasis on data engineering and we agree with the reviewer that this would need to be expressed in the text.

Author Changes:
See the changes proposed to Line 25 above.

Lines 380-386:
"A resilient modular monitoring system for urban environments has been developed to allow rapid new deployments with changes in infrastructure and network technology with a diverse set of field instruments being deployed during observation campaigns. The implementation primarily uses: freely-available software tools, established services for storing research data, and community adopted conventions. The system has to date been employed in several cities and different countries simultaneously. Our use cases not only involve research data products but also urban hydrometeorological services that reach the users – government officials, modeling teams and the public – in near real-time through the implementation of FAIR principles."

Reviewer Comment:
As mentioned, this project [h]as great potential to impact the observational science community with updated practices, architectures, demonstrated workflows, use cases, and lessons learned. Looking forward to seeing how it develops!

Author Response:
We thank the reviewer for those encouraging words.

Author Changes:
No change to the text.

Line 458:
Appendix A: Coordinate Systems

Reviewer Comment:
I found this entire paper to be valuable, but especially the detailed information in Appendix A. The use of absolute and relative coordinate systems and the concepts of documenting sensor deployment environments for urban zones should prove to be very useful for the observational community as these kinds of standards evolve.

Author Response:
We thank the reviewer for highlighting Appendix A. We hope to constructively contribute to communities working with observations in the urban zones, as more thought goes into the implementation and documentation of data systems.

Author Changes:
No change to the text.

[1] Scherer, D., Fehrenbach, U., Grassmann, T., Holtmann, A., Meier, F., Scherber, K., Pavlik, D., Hoehne, T., Kanani-Sühring, F., Maronga, B., Ament, F., Banzhaf, S., Langer, I., Halbig, G., Kohler, M., Queck, R., Stratbücker, S., Winkler, M., Wegener, R., and Zeeman, M.: [UC]2 Data Standard "Urban Climate under Change", Online, https://uc2-program.org/sites/default/files/inline-files/uc2_data_standard_0.pdf, version 1.5.2, 2022.