GAIA-CLIM Report / Deliverable D6.1

Gap Analysis for Integrated Atmospheric ECV CLImate Monitoring:

Report on results of user survey

A Horizon 2020 project; Grant agreement: 640276

Date: 31 July 2015

Lead Beneficiary: FMI

Nature: R

Dissemination level: PU

Work-package	WP 6 (Impact, outreach and dissemination)
Deliverable	D6.1
Title	Report on results of user survey
Nature	R
Dissemination	PU
Lead Beneficiary	Finnish Meteorological Institute, Helsinki
Date	31 July 2015
Status	Final
Authors	H. M. Mäkelä (FMI), J. Schultz (EUMETSAT)
Editors	Peter Thorne (NUIM), Anna Christina Mikalsen (NERSC)
Reviewers	M. De Mazière
Contacts	hanna.makela@fmi.fi, gerrit.leeuw@fmi.fi
URL	http://www.gaia-clim.eu/

This document has been produced in the context of the GAIA-CLIM project. The research leading to these results has received funding from the European Union's Horizon 2020 Programme under grant agreement n° 640276. All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors' view.

Contents

1 Introduction

GAIA-CLIM (Gap Analysis for Integrated Atmospheric ECV CLImate Monitoring) is one of the first H2020 space program projects and started in March 2015. The project is concerned with using ground-based, balloon-borne, and aircraft data to provide a sub-orbital cal/val capability for satellite data products. As the project is keen on interacting with data users and providers, a user requirements survey was undertaken at project inception. The objective of the survey was to help prioritize and shape the development of all aspects of the GAIA-CLIM project, but in particular, how data and information are to be served through a web-based facility, called the 'Virtual Observatory'. It was also timed such as to provide external user input to the first version of the Gap Analysis and Impacts Document (GAID). As the survey was carried out early in the project, it provides a baseline set of user expectations, as well as helping in identifying an initial group of core users who may be invited to future workshops and to test the tools developed within the project. Users who expressed a willingness to be contacted will be kept appraised of project progress and asked to provide appropriate input at later stages in the project lifecycle.

2 Survey design and dissemination

2.1 Survey design

The survey was composed of 15 questions concerning the i) respondent's background, ii) use of satellite, ground-based remote sensing, balloon-borne observations and aircraft observations data, iii) uncertainties, iv) data availability, format and tools, and v) the respondent's willingness to take part in the forthcoming GAIA-CLIM workshops (Table 1). The initial survey design was discussed in a breakout group and then presented to the whole group of partners at the Kick-off meeting in Matera, Italy. Subsequent to this, the underlying Work Packages provided input before a final version was agreed by all the Work Package 6 (outreach and dissemination) participants. The survey was intended to take respondents at most 10 minutes to complete in the hope of retaining those respondents who started the survey.

With the exception of Q13, each of the quantitative response questions was comprised of a varying number of predefined options from which a respondent could choose all that apply. In addition, the respondents could express their own alternatives under the option "Other, please specify" in each question. The complete survey with all its options is presented in Appendix A.

Concerning the Virtual Observatory (VO), questions in field 2 address the application areas of the VO, and questions in field 3 address which uncertainty information should be presented by the VO including what guidance is needed to understand the provided information. Finally, questions in field 4 are addressing the technical design of the VO, and questions in field 5 are searching for beta users.

Table 1. The 15 user requirement's survey questions divided into five survey fields.

1. RESPONDENT'S BACKGROUND (2 questions)	2. USE OF SATELLITE, GROUND-BASED REMOTE SENSING, BALLOON-BORNE OBSERVATIONS AND AIRCRAFT OBSERVATIONS (2 questions)	3. UNCERTAINTIES (3 questions)	4. DATA AVAILABILITY, FORMAT AND TOOLS (4 questions)	5. GAIA-CLIM WORKSHOPS WITH EXTERNAL USERS (4 questions)
Q1. What is your sector of work?	**Q3**. Are you a user of satellite data? If so how do you use it, and if not why don't you?	**Q5.** What is your level of expertise in using information about observational uncertainties related to satellite, ground-based, balloon-borne or aircraft data?	**Q8.** What timeliness of the observational data do you need for your intended applications?	**Q12.** Would you be interested in taking part in GAIA-CLIM workshops?
Q2. What is your principal field of work?	**Q4.** Are you a user of data arising from ground-based remote sensing, balloon-borne observations or aircraft observations? If so how do you use it, and if not why don't you?	**Q6.** Do you use uncertainty estimates?	**Q9.** GAIA-CLIM is planning to distribute all its data in NetCDF4 format containing Climate and Forecast compliant metadata. What is your preference for NetCDF4?	**Q13.** Do you have any suggestions for topics to be addressed in more detail during the workshops within the GAIA-CLIM project?
		Q7. Which guidance on how to utilize observational uncertainty information related to satellite and sub-orbital data would be valuable to you?	**Q10.** What kind of tools do you expect to be available on the Virtual Observatory?	**Q14.** Would you be interested in testing the tools developed in GAIA-CLIM?
			Q11. Which existing tools do you find useful to compare data sets?	**Q15**. If you wish you can leave your contact info here.

2.2 Survey dissemination and response rate

In order to collect feedback from the most relevant respondents, the survey was announced through the following communities in the Earth observation field:

- The World Climate Research Programme (WCRP) who highlighted the survey through e-zine and on their website;

- The Copernicus Climate Change Service (C3S) and the Copernicus Atmosphere Monitoring Service (CAMS) (both under ECMWF);
- EUMETSAT Satellite Application Facility (SAF) leads;
- ESA Climate Change Initiative (CCI) user community CMUG (the Climate Modelling User Group), both through an email and in person when presenting GAIA-CLIM at their coordination meeting held at SMHI;
- The International Ozone Commission (IO3C);
- The Network for the Detection of Atmospheric Composition Change (NDACC) Steering Committee members and UV/visible Working Group members;
- The Quality Assurance for Essential Climate Variables (QA4ECV) project;
- The Fidelity and uncertainty in climate data records from Earth Observations (FIDUCEO) project;
- The Advances in Atmospheric Science and Applications (ESA ATMOS) conference;
- CLIMLIST (an email distribution list for climatologists that consists of over 4,000 subscribers; Appendix B);
- Personal invites to the first GAIA-CLIM User Workshop
- Everyone in the GAIA-CLIM consortium (all participants and the advisory panel).
- Twitter distribution through several accounts

Through these channels, the survey invitation to participate will have reached at least 5,000 unique experts in this field. Although some individuals will have received the request several times the strategy was to deliberately minimize the number of such occurrences, which at least anecdotally can negatively impact perceptions of the requestors. The survey was released on 24 April and closed on 19 June, thus it was open for seven and a half weeks. Despite the substantial efforts made to collect input, the response rate remained fairly low. Altogether, 77 individual respondents completed the survey (Fig. 1a). The number of respondents grew quite smoothly during the time the responses were collected. The low response rate raises questions viz. “survey fatigue”, which may be relevant when considering such efforts in other projects in future.

Throughout the survey, almost all questions were answered by almost all the respondents (Fig. 1b). The exceptions were questions Q11 (about the existing tools) with 29 respondents, Q13 (free suggestions for workshop topics) with 10 respondents, and Q15 (contact information) with 37 respondents. For all the other questions, the number of responses varied from 68 to 77.

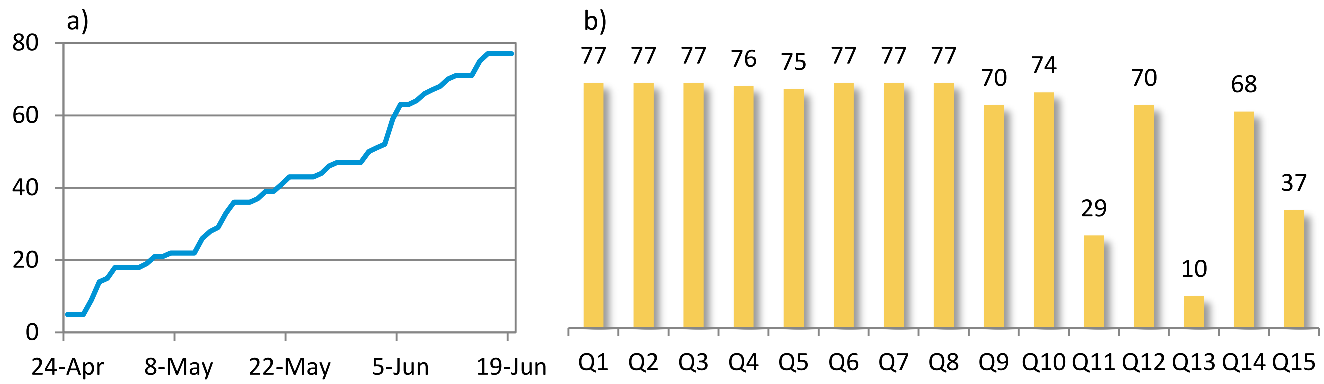


Figure 1. a) The number of respondents versus time. b) The number of respondents per question. Altogether 77 individual respondents took the survey.

2.3 Robustness of the survey

The low response rate raises a question as to representativeness and hence utility of the responses received. This relates to, i) whether the sample size is sufficient to make reasonable inferences and ii) whether the responses are truly representative of the broader user base.

The set of distribution choices was driven by a desire for a representative sample which is not skewed to any one sector or application area. To this end, the deliberate decision was made not to engage the reanalysis user base in particular, as had been done for a survey within the CORE-CLIMAX project, because that may have led to a larger but less representative set of responses. The user responses as to sector suggest that a balanced set of users answered the survey (Section 3.1), but this leaves open the question of how robust the responses may be, given that the sample size falls somewhat short of the 100-150 commonly seen as a reasonable minimum sample size to make robust inferences.

For a 95% confidence level (which means that there is a 5% chance that the survey sample results differ from the true population average), a good estimate of the margin of error (or confidence interval) is given by $1/\sqrt{N}$, where N is the sample size. With a sample size of 77 this leads to a margin of error of ±11.4%. Even though some tens of respondents more would have made the survey results more robust to draw conclusions, already the sample size of 77 can give some pointers to the opinions of the user base and for the development of the Virtual Observatory. In practice, the higher the consensus among the respondents regarding a certain survey question, the higher the chance that this is also the view of the broader user base. If the survey sample is skewed highly one way or the other the whole population probably is, too. However, in cases where the distribution of responses goes approximately to 50% (half of the respondents answered one way and the other half in another way), with margin of error around ±10% one cannot judge which is the dominant opinion, i.e., whether the votes goes 40%-60%, 60%-40%, or something in between.

To study the robustness of the survey responses the sample of 77 respondents was divided into two smaller subgroups by picking every other respondent in one group and every other in another group (in the chronological order of the responses). Statistical tests (namely the chi-squared test) applied then to the subgroups showed that the responses of these two subsamples were not dependent on the group. This gives confidence as regards the representativeness of the sample. It seems that no single group is dominating the results and that similar results would have been achieved even with a smaller amount of respondents.

3 Results of the survey

3.1 Respondent's background

Q1: "What is your sector of work?"
Q2: "What is your principal field of work?"

The majority, over 75 %, of the respondents were from the public sector R&D (research and development) or from academia (Fig. 2). The rest divided into private sector (10 %), other public sector and satellite agency (5 % each) and other (1 %). Four respondents chose more than one sector of work the combinations being: public R&D and academia (2 respondents), private and academia (1 respondent), and public R&D and other public sector (1 respondent). The only respondent that chose "Other, please specify" defined their sector of work "Private but research".

The respondents work mostly with "Climate" (60 % of the respondents) (Fig. 3). On average, the respondents picked 1.2 different fields of work. 17 respondents left a free form answer for their field of work (Table 2). The most popular were "Atmospheric composition" (5 respondents) and "Hydrology & Oceanography" (4 respondents).

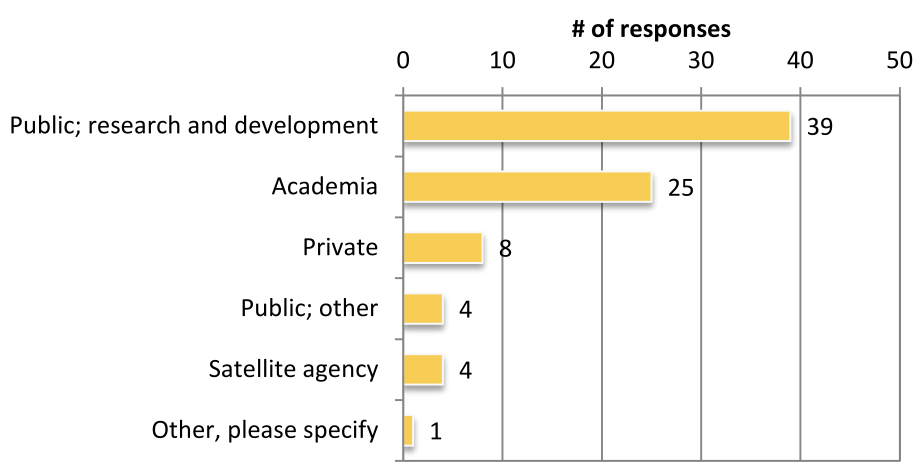


Figure 2. The respondents' sector of work (Q1). All 77 respondents answered this question, from which 4 respondents picked two options.

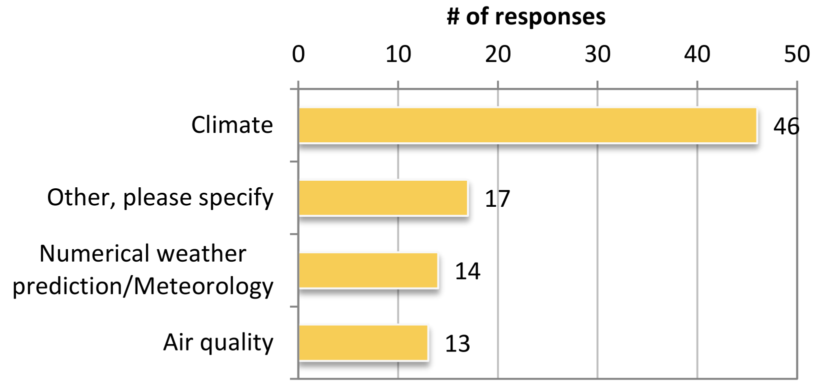


Figure 3. The respondents' field of work (Q2). All 77 respondents answered this question. Altogether 90 responses were given.

Table 2. The respondents' free form answers for the field of work (Q2). The number of respondents giving each field of work is given in the second column.

"Other, please specify"	# of respondents
Atmospheric composition	5
Hydrology & oceanography	4
Atmospheric chemistry	1
Atmospheric dynamics	1
Climate prediction	1
Data quality of the remote sensed observations	1
Development of software tools and services for earth observation	1
Ground-based remote sensing	1
Measurement and interpretation of solar UV radiation	1
Reanalysis	1

3.2 Use of satellite, ground-based remote sensing, balloon-borne observations and aircraft observations

Q3: "Are you a user of satellite data? If so, how do you use it, and if not why don't you? (please, choose all that apply)"
Q4: "Are you a user of data arising from ground-based remote sensing, balloon-borne observations or aircraft observations? If so how do you use it, and if not why don't you? (please, choose all that apply)"

Almost all respondents were users of satellite data (75 respondents, 97 %) (Fig. 4). Four of the respondents stated that they do not use satellite data; however, two of these picked also the "Yes" option indicating that they use the data but have issues with it ("it is too complicated to use", "I cannot find it" and/or "I cannot find the documentation").

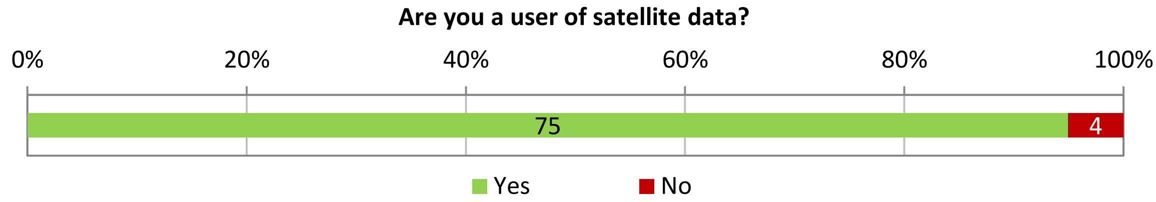


Figure 4. The use of satellite data (Q3). All 77 respondents answered this question.

The most common purpose for using satellite data was for model evaluation (42 respondents, 55 %), climate analysis (39, 51 %), and quality assessment of ground-based data (32, 42 %) (Fig. 5). The option "Other, please specify" received 21 responses (27 %) (Table 3). On average, the respondents picked 2.3 options for the use of satellite data.

The most common reason for not using satellite data was that it is too complicated to use (Fig. 5). There were also one or two respondents who could not find the data or the documentation, or who deemed that the data format or tools are not suitable. On average, those respondents who stated they did not use satellite data picked 2.5 options for their not using satellite data. Options "I do not need it", "I do not have access to it", "There are no suitable datasets for my applications", or "I do not trust it" were not chosen at all.

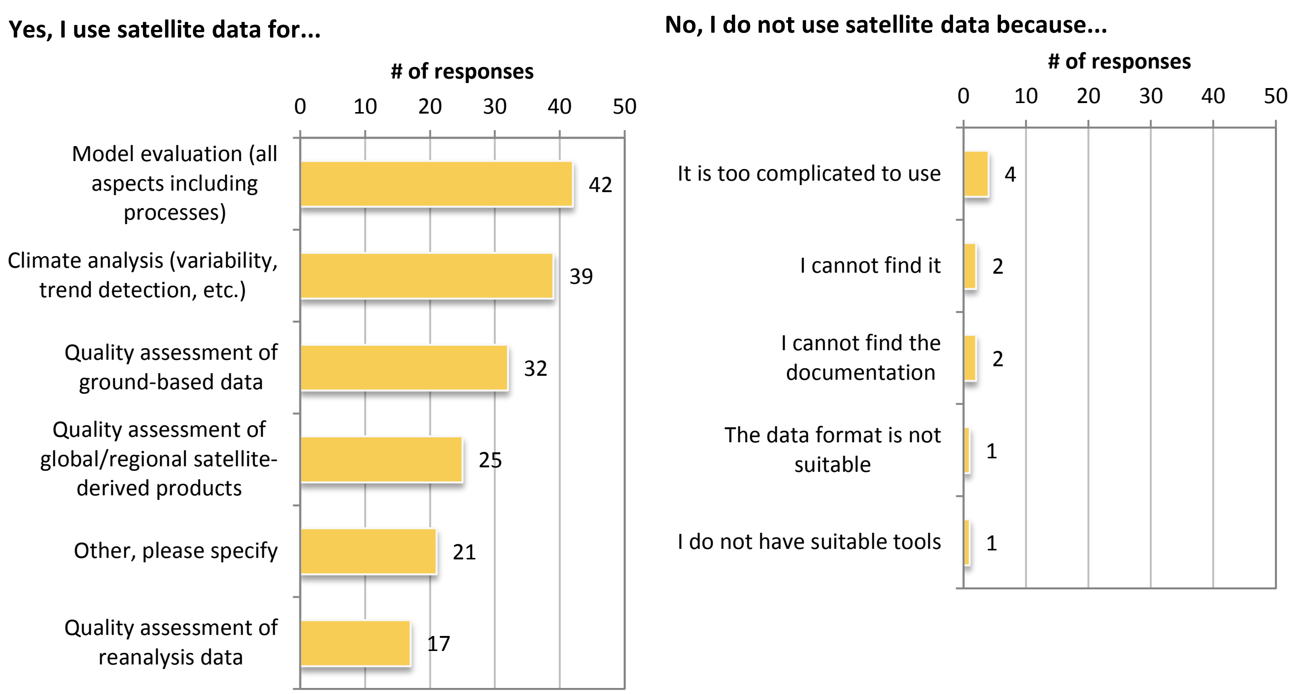


Figure 5. Reasons given for the use (left) and non-use (right) of satellite data (Q3). All 77 respondents answered this question.

Table 3. The respondents' free form answers for the use of satellite data (Q3). The number of respondents giving each purpose of use is given in the second column. Similar responses have been grouped.

"Other, please specify"	# of respondents
Data assimilation	5
Quality assessment of satellite data itself	3
Context for aircraft/ground data e.g. to confirm back-trajectories	2
Educating forecasters	2
1) Generation of meteorological products 2) As a model input	1
As complement to data of ground-based measurements	1
Case study analyses	1
Critical examinations of physical/statistical interpretation	1
Determination of sea ice conditions	1
Global assimilation for monitoring (and forecasting) of atmospheric composition	1
I overlay it with other remote sensing data	1
Planetary sciences	1
Production of satellite data for climate and other applications	1

Data arising from ground-based remote sensing, balloon-borne observations, or aircraft observations was used by 69 respondents (91 %) (Fig. 6). Seven respondents stated that they do not use these data. One respondent uses it, but picked also the "No" options "it is too complicated to use" and "I do not have suitable tools".

The most common purposes for the use of data arising from sub-orbital measurements were model evaluation (42 respondents, 55 %), quality assessment of satellite retrievals (36, 47 %), climate analysis (33, 43 %) and comparison to other in-situ measurements (32, 43 %) (Fig. 7). The three most common purposes are the same as were given for satellite data. On average, the respondents picked 3.3 options for the use of the data. The option "Other, please specify" received seven responses (9 %) (Table 4). The most common reason for not using the data was that it is not needed (Fig. 7).

There were also one or two respondents who found that there are no suitable datasets or tools, or that the data is too complicated to use. In the free-form comments one respondent stated that they cannot use the data because of missing observations. On average, the respondents picked 1.1 options for not using these data. Options "I do not have access to it", "I cannot find it", "I cannot find the documentation", "The data format is not suitable", or "I do not trust it" were not chosen at all.

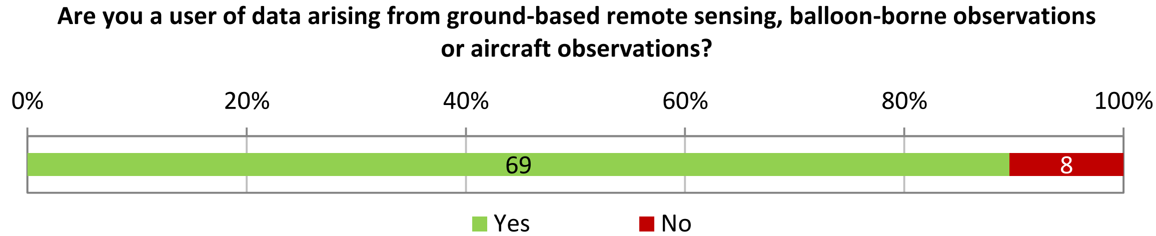
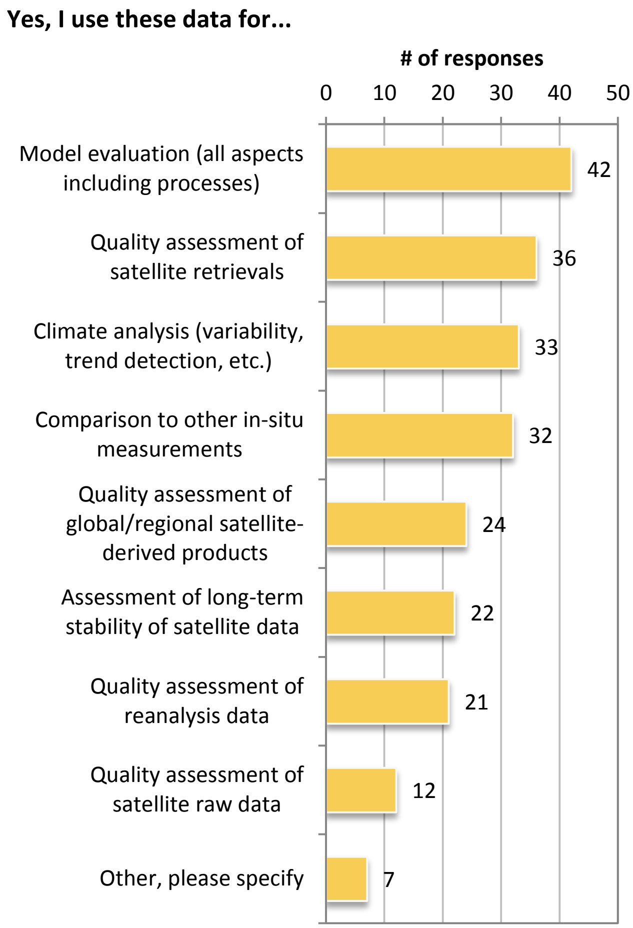


Figure 6. The use data arising from ground-based remote sensing, balloon-borne observations, or aircraft observations (Q4). 76 respondents answered this question.



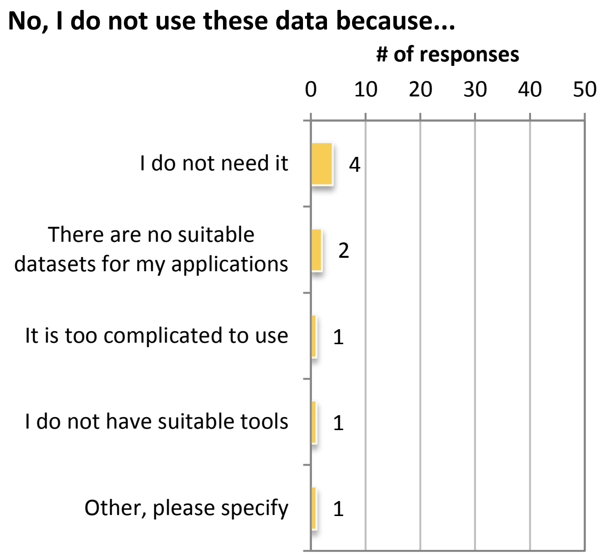


Figure 7. Reasons for the use (left) and non-use (right) of data arising from ground-based remote sensing, balloon-borne observations, and aircraft observations (Q4). 76 respondents answered this question.

Table 4. The respondents' free form answers for the use of data arising from ground-based remote sensing, balloon-borne observations and aircraft observations (Q4). The number of respondents giving each purpose of use is given in the second column. Similar responses have been grouped.

"Other, please specify"	# of respondents
Assimilation, now-casting	2
1) As a model input 2) As a source of data in its own right.	1
Critical examinations of physical/statistical interpretation	1
Quality assessments of radiosonde data, will also be doing this for airborne data	1
Study of atmospheric processes	1
Weather forecasting, severe events	1

3.3 Uncertainties

Q5: "What is your level of expertise in using information about observational uncertainties related to satellite, ground-based, balloon-borne or aircraft data? (please, choose all that apply)"
Q6: "Do you use uncertainty estimates? (please, choose all that apply)"
Q7: "Which guidance on how to utilize observational uncertainty information related to satellite and sub-orbital data would be valuable to you? (please, choose all that apply)"

The respondents' level of expertise in using information about the observational uncertainties related to satellite, ground-based, balloon-borne, or aircraft data was asked through a scale from one (1=little expertise) to five (5=professional) (Figure 8). 75 respondents answered this question, however, none of those for all the parts. This likely reflects the complexity of the use of uncertainty information and issues surrounding the ability of participants to self-assess their expertise. Overall, the respondents felt most professional in using gross uncertainties (the average level of expertise being 3.7), and uncertainties arising from systematic (3.6) or random (3.5) effects of the measurements (Fig. 8). The respondents had the least expertise in using fully traceable uncertainty estimates (the average level of expertise being 2.8), and full uncertainty budget per observation system (3.0). Under the "Other, please specify" one respondent stated that the level of expertise depends on the observational system and application.

For the question "Do you use uncertainty estimates?" (Q6) 67 respondents picked "Yes" (85 %) and 12 opted "No" (15 %). Two of the respondents belonged to both groups, having opted both "Yes" and "No". The most common reason for not using these data was that it is considered too complicated to use (5 respondents, 6 %), while one does not have access to it (3 respondents, 4 %) and one cannot find it (3 respondents, 4 %) (Fig. 9) were next most frequent.

As for guidance on how to utilize observational uncertainty information the respondents would most preferably use classical user guide document (the average score for preference was 4.3 at a scale from one (1=least preferred) to five (5=most preferred)) (Fig. 10). Less preferred options would be a list of peer reviewed papers (3.4), a help desk providing support on request (3.2), and an online training course (3.1). Four respondents left their own suggestions under the option "Other, please specify" (Table 5).

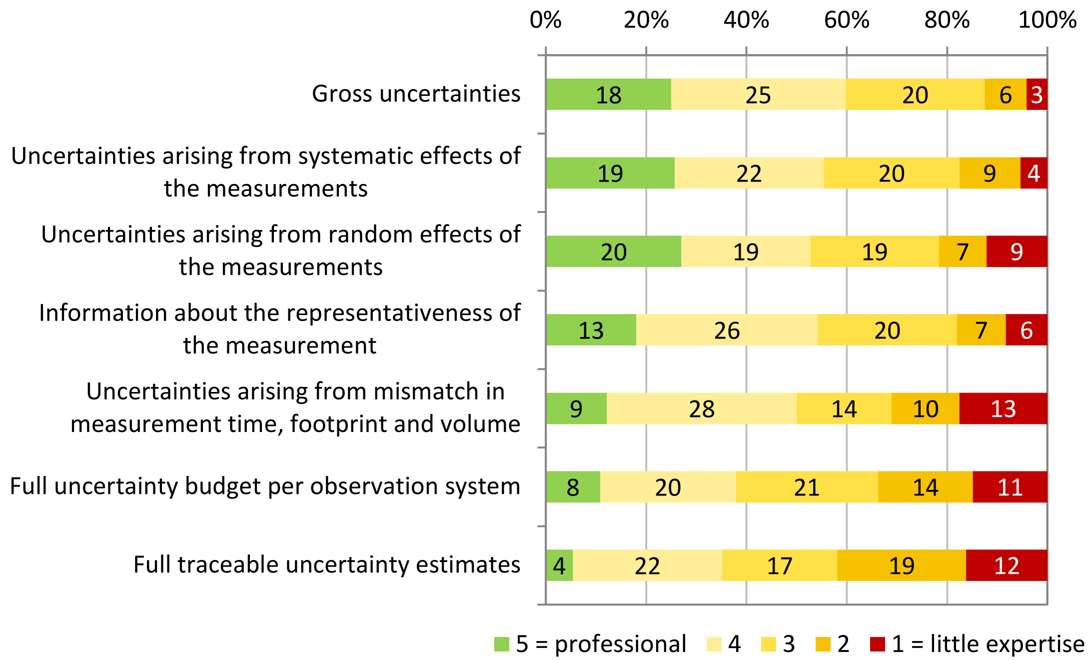


Figure 8. The respondents' level of expertise of in using information about observational uncertainties related to satellite, ground-based, balloon-borne or aircraft data (Q5). The numbers show the number of the respondents in each category. (In all cases the categories are ranked left to right with leftmost being score 5 (professional) and rightmost being 1 (little expertise) and numbers given.)

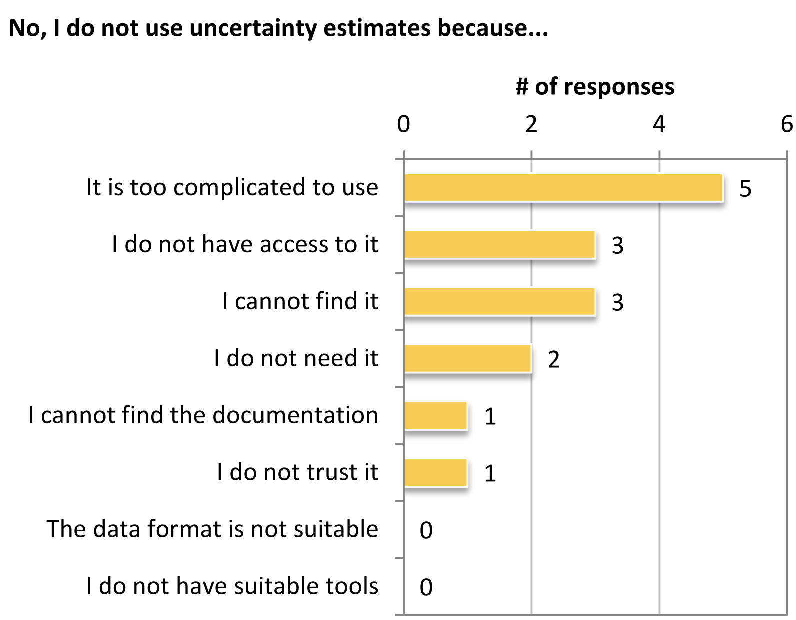


Figure 9. Reason for not using uncertainty estimates (Q6). All 77 respondents answered this question.

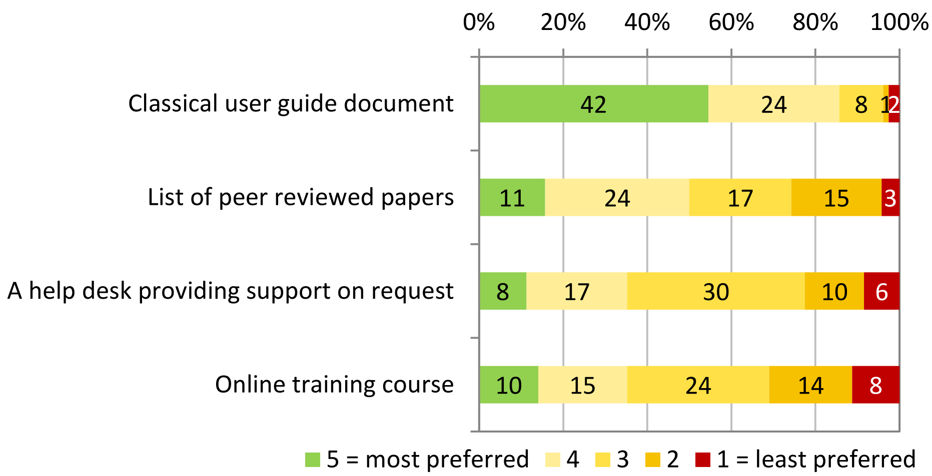


Figure 10. The respondents' preference for different guidance on how to utilize observational uncertainty information related to satellite and sub-orbital data (Q7). The figures show the number of the respondents in each category. All 77 respondents answered this question. (In all cases the categories are ranked left to right with leftmost being score 5 (most preferred) and rightmost being 1 (least preferred) and numbers given.)

Table 5. The respondents' free form answers for the desired guidance on how to use observational uncertainty information (Q7). The number of respondents giving each suggestion is given in the second column.

"Other, please specify"	# of respondents
A workshop run by instrument experts	1
A web page showing examples	1
A concise traceable document providing key information	1
Ensemble of observations	1

3.4 Data availability, format and tools

Q8: "What timeliness of the observational data do you need for your intended applications? (please, choose all that apply)"
Q9: "GAIA-CLIM is planning to distribute all its data in NetCDF4 format containing Climate and Forecast compliant metadata. What is your preference for NetCDF4? Other format preferences, please specify."
Q10: "What kind of tools do you expect to be available on the Virtual Observatory? (please, choose all that apply)"
Q11: "Which existing tools do you find useful to compare data sets? (please, choose all that apply)"

Q8 enquired as to the desired timeliness of the observational data that the respondents need for their intended applications. Around 70 % of the respondents (55 respondents) favoured timeliness of offline (1 year) and/or delayed mode (1 month), and/or did not consider it to be important (Fig. 11). That leaves 22 respondents (29 %) who indicated that near real-time (NRT) timeliness, either alone or combined with some other option, would be needed. On average, the respondents picked 1.6 choices. The most popular choices were offline (17 respondents, 22 %), delayed mode (16 respondents, 21 %), and the combination of the two: delayed mode + offline (11 respondents, 14 %) (Table 6).

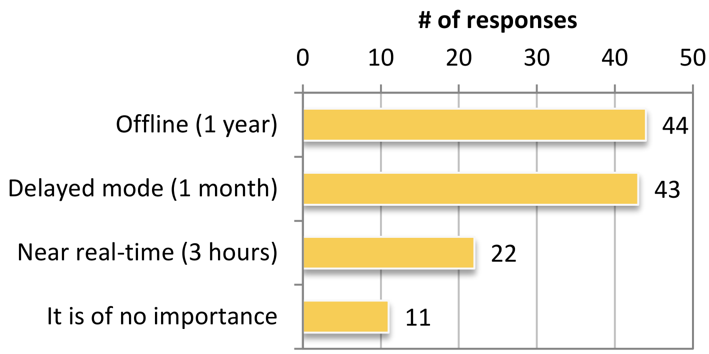


Figure 11. The respondents' desired timeliness of the observational data (Q8). All 77 respondents answered this question. Each respondent picked on average 1.6 choices.

Table 6. The respondents' desired timeliness of the observational data (Q9). The number of respondents giving each suggestion is given in the second column. All 77 respondents answered this question.

Timeliness	# of respondents
Offline	17
Delayed mode	16
Offline + delayed mode	11
Offline+ delayed mode + near-RT	9
No importance	7
Near-RT	5
Near-RT + delayed mode	4
Near-RT + offline	4
Offline + delayed mode + no importance	2
Delayed mode + no importance	1
Offline + no importance	1

GAIA-CLIM is planning to distribute all its data in CF-compliant NetCDF4 data format. When asking respondents preference for NetCDF4, a clear majority would be satisfied with it (Fig. 12). Only three respondents picked the two lowest preference categories. Their suggestions for suitable data format were HDF5 and ASCII. In total 22 respondents left suggestions for desired data format (Table 7). The most frequently suggested additional data formats were HDF5 (9 respondents), HDF (4 respondents), which could be classed together as 13 respondents requiring a variant of HDF as frequently the version is not quoted by users, and ASCII (3 respondents; 4 if the user specifying flat ASCII is included).

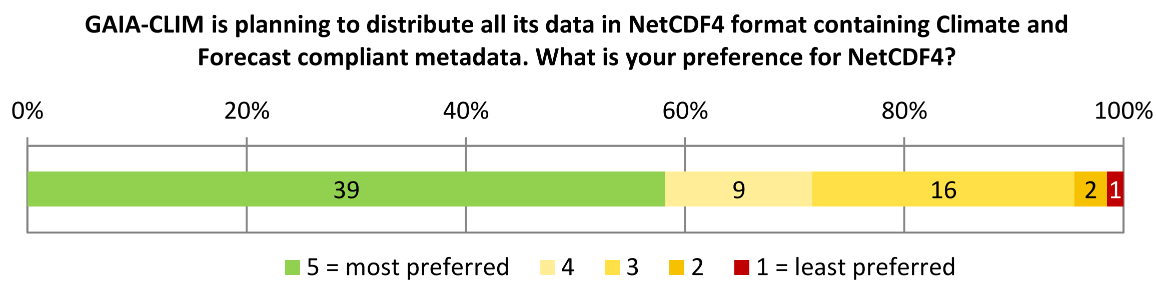


Figure 12. The respondents' preference for NetCDF4 data format (Q9). (The categories are ranked left to right with leftmost being score 5 (most preferred) and rightmost being 1 (least preferred) and numbers given.)

Table 7. The respondents' free form answers for the desired data format (Q9). The number of respondents suggesting each data format is given in the second column. In the third column an average of the preference (at a scale from one (1=least preferred) to five (5=most preferred)) of each suggested data format is given.

"Other format preferences, please specify"	# of respondents	Preference
HDF5	9	4.00
HDF	4	4.25
ASCII	3	4.33
ODB	2	4.00
BUFR	1	5.00
CSV	1	5.00
FITS	1	5.00
GRIB	1	5.00
NetCDF3	1	5.00
Flat ASCII	1	4.00
HDF4	1	4.00
NetCDF-CF	1	4.00
v4	1	4.00
ARTS-XML	1	2.00

Respondents' preference for desired tools to be available on the Virtual Observatory was ascertained through various prompted suggestions (Q10) (Figs. 13a-f). All the suggested tools were found to be more desired than not; the average preference score varied from 3.0 (smoothing of vertical resolution) to 4.5 (sub-setting by variable) at a scale from one (1=least preferred) to five (5=most preferred).

For data search and discovery, the possibility of sub-setting by variable, time and location were found to be very important (Fig. 13a). Over 80 % of the respondents placed those features in one of the two categories of highest preferences; the average preference scores varied between 4.4 and 4.5. Less desired sub-setting options were sub-setting by observing system (4.0) and by surface type (3.4).

For online data post-processing and data analysis, the respondents found most important to have a possibility of regridding data to a common resolution (3.8) and to be able to average data in space and/or time (3.6) (Figs. 13b and c).

The possibility to have a flexible forward modelling capability divided respondents roughly in half (Fig. 13d); however, the average preference for this feature remained positive (3.1).

For data extraction and visualization tools, the most desirable features were collocated data files (4.1) and the ability to produce geographical maps (4.0) (Figs. 13e and f).

Six respondents left also free form suggestions for desired tools (Table 8).

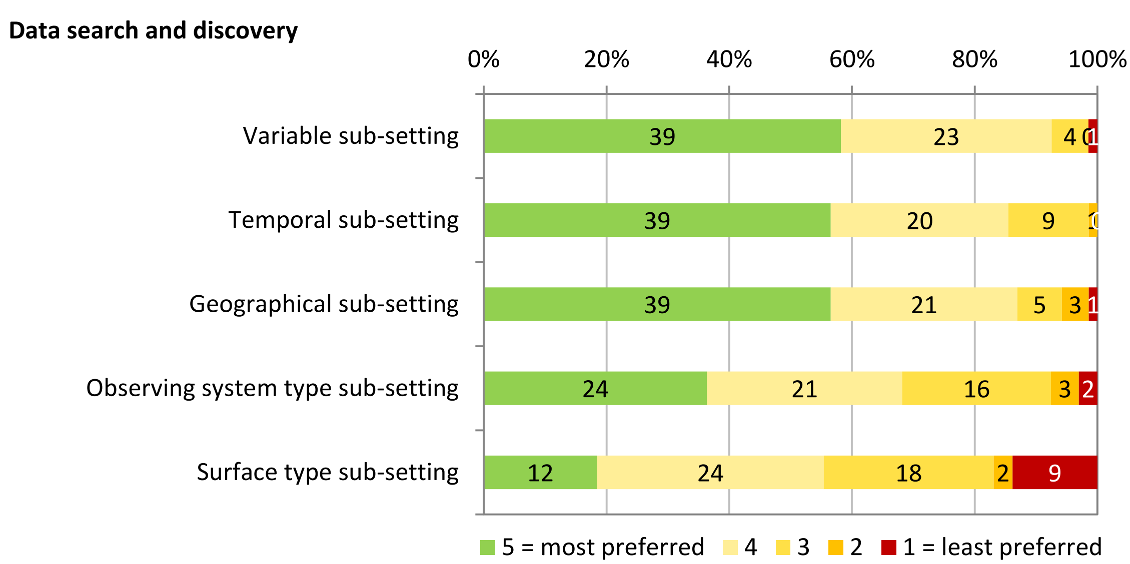
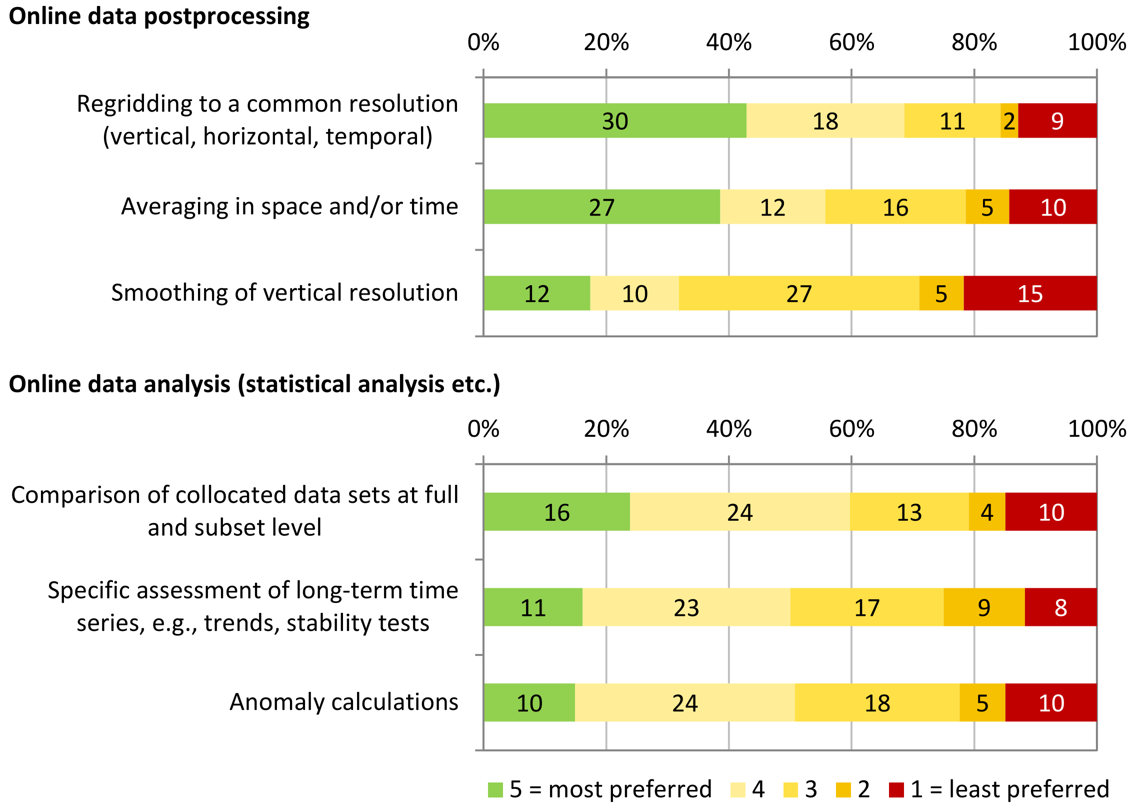


Figure 13a. The respondents' preference for different data search and discovery tools (Q10). The figures show the number of the respondents in each category. (In all cases the categories are ranked left to right with leftmost being score 5 (most preferred) and rightmost being 1 (least preferred) and numbers given. Variable sub-setting had no responses in category 2 and 1 in category 1; temporal sub-setting had 1 response in category 2 and none in category 1.)



Figures 13b and c. The respondents' preference for different online data post-processing and data-analysis tools (Q10). (In all cases the categories are ranked left to right with leftmost being score 5 (most preferred) and rightmost being 1 (least preferred) and numbers given.)

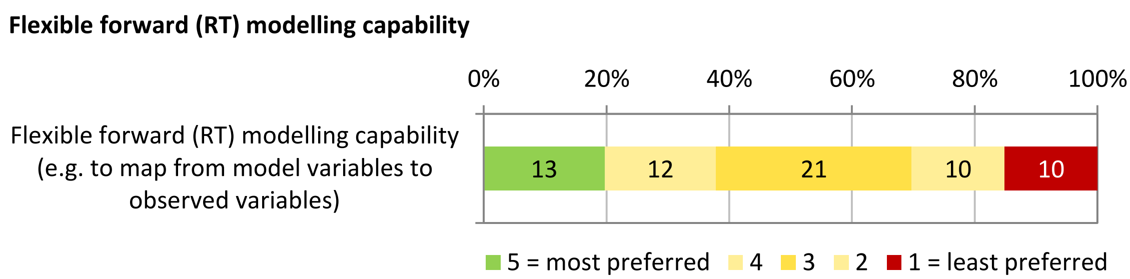
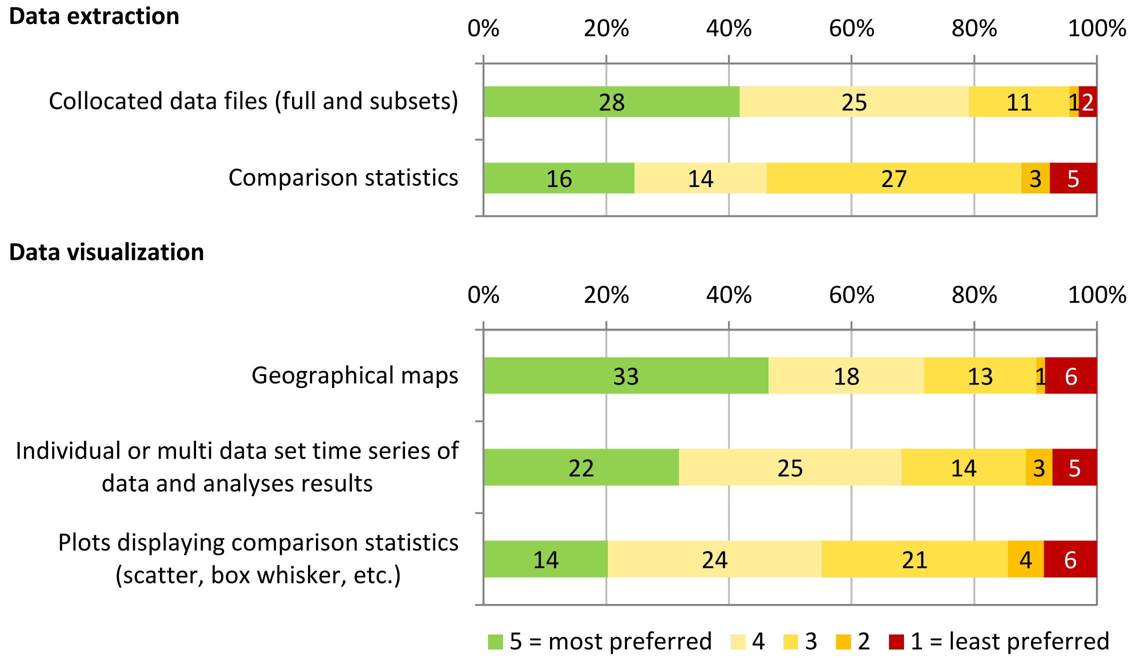


Figure 13d. The respondents' preference for flexible forward modelling capability (Q10). (In all cases the categories are ranked left to right with leftmost being score 5 (most preferred) and rightmost being 1 (least preferred) and numbers given.)



Figures 13e and f. The respondents' preference for different data extraction and data visualization tools (Q10). (In all cases the categories are ranked left to right with leftmost being score 5 (most preferred) and rightmost being 1 (least preferred) and numbers given.)

Table 8. The respondents' free form answers for the desired tools (Q10). The number of respondents suggesting each feature is given in the second column. Six individual respondents left free form answers for this question.

"Other, please specify"	# of respondents
Data search and discovery	
An API to get all metadata of all available products that match certain basic query criteria (measurement time, last metadata update time, location, instrument/measurement type)	1
A fast SQL interface	1
Online data post-processing	(none)
Online data analysis (statistical analysis, etc.)	(none)
Flexible forward (RT) modelling capability	(none)
Data extraction	(none)
Data visualization	(none)
Other, please specify	
All the above, but not performed by button pressing, but perhaps given a library, performed by a script that I can edit and test locally and/or submit online.	1
Ability to download data directly (i.e. direct download URLs)	1
FTP	1

We mostly tailor tools for our own applications and studies. Community developed open source tools are also applied. In this respect it is important that open formats are used and metadata standards followed.	1
Toolbox of scripts for data handling that can be downloaded and modified for personal use.	1

The respondents use currently various different tools to compare data sets (Q11); however, only 29 respondents answered this question. This suggests that the list of tools named was not regarded by the survey participants as complete, and that no leading, widely-used tool for the comparison of satellite and suborbital observations exists to date. Most of the respondents picked something else ("Other, please define", 16 respondents) than the predefined options: AEROCOM (8 respondents), ICARE (6), NORS Portal and NPROVS (3 respondents each) (Fig. 14). In Table 9 the respondents' free form answers for the used existing tools are listed. An option that was mentioned most often was "my own tools" or "institute's tools" (6 respondents).

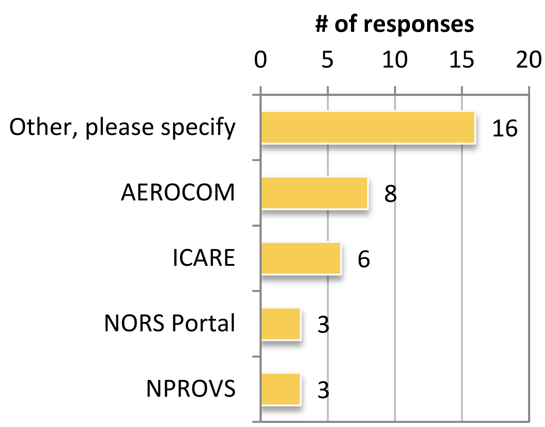


Figure 14. The existing tools used by the respondents (Q11). Only 29 respondents answered this question.

Table 9. Respondents' free form answers for the used existing tools (Q11). The number of respondents suggesting each tool is given in the second column.

"Other, please specify"	# of respondents
My/Institute's own	6
GIOVANNI (NASA)	2
NDACC	2
AERONET, African Climate Atlas, AVDC (NASA), CIS, Climate Data Operators, Igor Pro, programming tools, satellite data websites, TCCON, TEMIS	1

3.5 GAIA-CLIM Workshops with external users

Q12: "Would you be interested in taking part in GAIA-CLIM workshops? If you wish to get notifications about the workshops by email, please, tick yes and enter your contact info in question 15 below. The first workshop will be held on October 6, 2015 in Rome, Italy."
Q13: "Do you have any free suggestions for topics to be addressed in more detail during the workshops within the GAIA-CLIM project?"
Q14: "Would you be interested in testing the tools developed in GAIA-CLIM? If you wish to be contacted concerning the testing, please, tick yes and enter your contact info in question 15 below"
Q15: "If you wish you can leave your contact info here."

The last survey field enquired about the respondents' willingness to take part in GAIA-CLIM workshops or testing the tools developed in GAIA-CLIM (Fig. 15). 28 respondents (36 % out of all 77 respondents) indicated that they would be interested in GAIA-CLIM workshops. 21 of these left their name and email information so that they can be invited and informed about the workshops. A somewhat smaller group, 22 respondents (29 % out of all 77 respondents), responded that they would be interested in testing the tools developed in GAIA-CLIM. 20 of these left their contact information. There were 13 respondents that were interested both, in the workshops and in testing the tools.

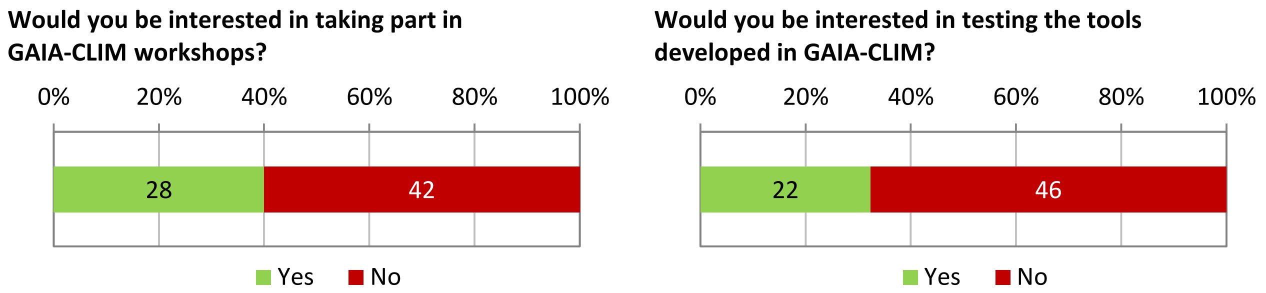


Figure 15. The respondents' willingness to take part in GAIA-CLIM workshops (left, Q12) and testing the tools developed in GAIA-CLIM (right, Q14). 70 respondents answered Q12 and 68 respondents answered Q14.

Ten respondents left suggestions for topics to be addressed in more detail during the workshops within the GAIA-CLIM project (Q13). The list of the suggestions is given below. Each of the suggestions in the list was given by one respondent.

- Topical campaigns for specific research themes
- Potential inter-project synergies between FIDUCEO and GAIA-CLIM
- Statistical methods to incorporate measurement uncertainties in the calculation of spatial and temporal means
- Address the issue of disseminating level 1 products and its uncertainty instead of satellite retrievals

- Methods for relating global surface measurement networks to satellite observations
- Forward modelling - and how to do it properly
- Long-term trends; data set stability; averages, e.g., monthly zonal means
- The most relevant model of propagation of uncertainties/errors when performing comparisons between two measurements of the same object with two different methods
- Ways to handle space/time collocation mismatches
- Be clear about definitions of uncertainty (bias, accuracy, stability ...)

4 Consequences for definition of Virtual Observatory

As briefly outlined in section 2 the survey was designed to support the definition of the functionality and partly the content of the Virtual Observatory. The expectation was that the information from the survey can be directly used to draft a user requirements document that could be further discussed with those respondents who expressed an interest in continued engagement. Ultimately, the user requirements should lead to system requirements that will determine the logical and physical design of the VO. The following sections provide an initial interpretation of the survey answers with respect to what they mean for the design and functionality of the VO.

4.1 Respondent's background

The fact that about 60% of the responders named climate as their major field of work indicates that the content of the VO in terms of data sets should give priority to long time series whenever they are available from sub-orbital reference-quality measurement systems and satellite data. As interest in climate also implies an importance of process level studies a second priority may be to consider satellite data with highest accuracy. The two secondary groups of interest on numerical weather prediction and air quality indicate a need for some of the content of the VO towards atmospheric state and composition data. This gives some prioritization as to initial case studies and products to offer during the development of the VO. However, it should be stressed that overall design and capability of the VO will be kept general to allow for future extensions to other Earth System domains, in particular for the Copernicus Climate Change Service.

4.2 Use of satellite, ground-based remote sensing, balloon-borne observations and aircraft observations

As for both the satellite and sub-orbital measurements the most frequent intended application is model evaluation the VO design should think about how to link the satellite and sub-orbital data with climate model intercomparison activities. One potential approach would be to include extracted CMIP5 model results in addition to the reanalysis data. As this is not foreseen in the current project plan WP5 needs further discussion of what needs to be technically done and what it may cost before a decision is taken as to whether this extension to functionality is viable. As the second and third ranked application areas in both cases are climate analysis and quality assurance the survey has highlighted the major usage is climate-based applications that will receive priority in being served by the VO (Section 4.1). All three major envisaged application areas match the initial project consortium expectation and were already envisaged in the planning of the VO.

The most interesting information on the use of satellite data for the VO is the relatively large number of users who expressed a need to use these data for the quality assessment of ground-based data which seems mostly motivated by the spatiotemporal homogeneity of most satellite products, in particular at lowest levels of processing, e.g., radiances. For the VO it may indicate that functionality of bringing ground-based data to observation space by applying forward radiative transfer modeling (which is well-posed unlike the opposite approach) enabling comparisons at radiance level is attractive. However, the answers received for the specific question on the forward radiative transfer capability were only seen by about 25% as a preferred functionality. The reasons for this apparent dichotomy are unclear and some interaction with willing users on this topic might be necessary to clarify. However, the possibility to undertake comparisons in radiance space remains a planned activity regardless.

The reasons mentioned by the 3% of respondents not using satellite data serves to keep the project aware of the importance of easy access and understandable documentation in building the VO.

4.3 Uncertainties

Answers on the uncertainty field related questions are not so easy to interpret in their value for the design of the VO. The distribution of answers clearly shows that the knowledge about uncertainty measures is rather low if questions move away from well-known statistical terms. In particular the concept of traceability seems not very well known among many of the survey participants. For the design of the VO this means that with each uncertainty measure included from developments in other work packages we need to have a very good and relatively easy understandable explanation what it means for different application areas. The preferred way for communicating this information seems to be an elaborated dedicated user guide document which represents a clear user requirement. This can also satisfy the second most requested option of the list of peer reviewed papers. It appears that a dedicated help-desk is not seen, at least at this point, by respondents as a priority facility to pursue.

4.4 Data availability, format and tools

Answers to the timeliness question are fully consistent with the work field (climate) and the mentioned major intended applications. The majority of respondents have no strong timeliness requirement, which means that for the majority of likely users the VO can be successfully working in an offline delayed service mode. However, a quarter of respondent expects near real-time availability which is certainly linked with the fact that 14 respondents come from the numerical weather prediction area. It is important information for the VO in the sense that the design needs to foresee an eventual operational near real time capability even if implemented after the end of the project. It may also trigger the need of very careful data quality flagging for quasi real-time data input streams and the need for reprocessing of data collocations for users of offline data which may have higher demands on the quality.

The data format question provides a very clear answer to the design of the VO, that is, to realise NetCDF formatted data extraction with maybe one alternative format, e.g., HDF5. Interestingly, there is only one request for the WMO standard BUFR format. The answers on the needed tools allow a clear prioritization for the implementation of the VO. In addition, the very small number of additional proposals might be interpreted in a way that the initially proposed tools match user needs quite well. Very interesting is that only one respondent has the wish for a software toolbox compared to online processing capability. This is important to know for the VO because distributing software means much more maintenance work and interaction with specific users of the software which would most likely increase the cost for a later operational use of the VO.

The question about the existing tools for comparing data sets drew only 29 respondents indicating that the list of tools named was not regarded by the survey participants as complete, and that no leading, widely-used tool for the comparison of satellite and suborbital observations exists to date. This is positive in that there is a clear need for such a facility as envisaged in the VO but also negative in the sense that whatever tool will be provided it will not necessarily match all individual needs. However, those needs can be served through the data extraction functionality and the use of metadata complying with international standards.

4.5 Next steps

Despite the relatively small number of responses the survey was successful in providing specific information for defining more formal user requirements for the VO which will be the next step forward from the survey. Of some concern is the relatively low willingness of the respondents to engage in the development of the VO. Nevertheless, it is planned to circulate the user requirements to those who have left their contact details. In addition, it can be tried to figure out key people from the most mentioned application areas that can be asked to review a user requirements document for the VO. The answers to the user requirements will inform a technical design document that lists all the functionalities the VO will have and how they will be realized. This includes also a compliance matrix with respect to the user requirements that will enable the project to trace back the implementation to the user survey.

5 Consequences for the Gap Analysis and Impacts Document

Although the survey was primarily aimed at informing the development of the Virtual Observatory and therefore touched only upon a subset of the totality of the GAIA-CLIM project objectives, it did serve to highlight a number of points that are relevant to the identification of gaps and assessment of their impacts. The most salient of these are highlighted here as potential input to the first version of the Gap Analysis and Impacts Document (GAID).

The survey was distributed to a broad range of potential end-users of GAIA-CLIM outcomes via a combination of group emails to relevant groups, web-based advertising, twitter, personal invitations to selected high-priority targets and presentations at relevant conferences. Despite this, the participation rate was low. As stated in Section 2, this may simply reflect a degree of "survey fatigue" on the part of the user community. It is, of course, impossible to know for certain why non-responders did not respond. A possible explanation is that it points to a gap in education and knowledge about how sub-orbital and/or satellite data can be used for both, scientific and practical applications, which is difficult to assess by the users themselves ("You cannot know what you do not know."). This gap in knowledge will partly be addressed if GAIA-CLIM is successful, but also highlights the need for a strong outreach component to GAIA-CLIM and the Virtual Observatory.

- ➔ The relative lack of response rate and its implications

The responses to Q11 highlighted that very few of the listed pre-existing match-up database facilities, such as NORS, ICARE and NPROVS, are currently used by those who participated. Instead most users who responded registered a preference for use of their own tools or these in combination with existing facilities. This highlights two substantive possible gaps. First, a gap in knowledge as to what interface capabilities users are looking for. If users had clearly defined an existing service that met their needs this would have provided strong development guidance to the Virtual Observatory. Secondly, a potential reluctance of users (always assuming the responders are representative) to take up new tools; unless they adequately replicate the functionality of their own tools, which they know and understand. The answers to Q10 become an important aspect of building a service that will be useable and used and represent a potential way that GAIA-CLIM can address this gap.

- ➔ The lack of broad-scale use of existing match-up facilities

Responses to Q8 highlighted a broad range of expertise in the use of uncertainty information. More than half of the participants rated their ability to use full uncertainty budget or traceable uncertainty estimates in the lowest three of five categories. This represents a gap if the intended GAIA-CLIM target data are to be used appropriately by end-users as a key aspect of GAIA-CLIM is the use of traceable uncertainty estimates to constrain the comparisons between satellite and sub-orbital measurements. The gap could be addressed through the preparation of educational materials, tools and the publication in the literature of a number of examples where the data and their uncertainties are used appropriately.

- ➔ The broad range of user familiarity with and use of uncertainties

6 Broader implications for GAIA-CLIM

The survey has served to help highlight user preferences for the Virtual Observatory (Section 4) and a number of potential gaps to be addressed in the GAID (Section 5). Beyond this, the relatively low response rate highlighted the need to have a strong and sustained outreach program focusing on user engagement and education. The design of the Virtual Observatory has implications for other Work Packages, which shall deliver as their outcomes inputs to the Virtual Observatory. Implications of the survey outcomes for these underlying work packages shall be addressed in upcoming meetings of the management group and general assembly. The most important fora, however, will be the series of user workshops. All individuals that indicated further interest received a personal invitation to the first user workshop, which will take place in early October.

7 Summary

The survey was conducted as an early deliverable of GAIA-CLIM to capture user expectations and engender interest in the project. The survey was advertised broadly. However, it only received a relatively low response rate. Because of the relatively small sample size the margin of error of the survey remained as high as ±11%. However, the respondents represented a broad range of user areas and according to a statistical tests the sample seemed to be robust in a way that similar results were achieved even with smaller subsamples. Overall, the survey was successful in providing specific information upon which to narrow down more formal user requirements for the Virtual Observatory which will be the next step forward from the survey.

Based on the survey the priorities of VO will be to present long time series (because majority of the responders were interested in climate), and to serving the application areas of model evaluation, climate analysis, and quality assessment of ground-based data/satellite retrievals. All these three application areas match the initial project consortium expectation and were already envisaged in the planning of the VO. Development of very good and easily understandable explanations especially regarding uncertainty measures is needed, and desired in the form of a classical user guide document. As the majority of the respondents has no strong timeliness requirement it means that the VO can be successfully working offline (in delayed data mode) in its initial implementation. However, especially those respondees concentrating on the numerical weather prediction application area expressed a desire for near real-time availability. This is important to keep in mind in the sense that the design needs to foresee this eventual capability, even if implemented after the end of the project.

The central gaps that were revealed by the survey and which will be potential input to the first version of the Gap Analysis and Impacts Document were:

i. the lack of response rate implicating a possible gap in education and knowledge about how sub-orbital and/or satellite data can be used for scientific and practical applications;

ii. the lack of broad-scale use of existing match-up facilities implicating a clear need for such a facility, but on the other hand a difficulty to define a functional one, and maybe also a potential reluctance of users to take up new tools;
iii. the broad range of user familiarity with and use of uncertainties implicating a need for education and capacity building in this area.

Appendix A. The GAIA-CLIM User Requirements Survey

GAIA-CLIM User Requirements Survey

Gap Analysis for Integrated Atmospheric ECV CLImate Monitoring

Respondent's background (1/5)

1. What is your sector of work?

- ☐ Private
- ☐ Public; research and development
- ☐ Public; other
- ☐ Academia
- ☐ Satellite agency
- ☐ Other, please specify

2. What is your principal field of work?

- ☐ Numerical weather prediction/Meteorology
- ☐ Air quality
- ☐ Climate
- ☐ Other, please specify

Use of satellite, ground-based remote sensing, balloon-borne observations and aircraft observations (2/5)

3. Are you a user of satellite data? If so how do you use it, and if not why don't you? (please, choose all that apply)

a. Yes, I use satellite data for

- ☐ Model evaluation (all aspects including processes)
- ☐ Quality assessment of reanalysis data
- ☐ Quality assessment of global/regional satellite-derived products
- ☐ Quality assessment of ground-based data
- ☐ Climate analysis (variability, trend detection, etc.)
- ☐ Other, please specify

b. No, I do not use satellite data because

- ☐ I do not need it
- ☐ I do not have access to it
- ☐ I cannot find it
- ☐ It is too complicated to use
- ☐ I cannot find the documentation
- ☐ The data format is not suitable
- ☐ I do not have suitable tools
- ☐ There are no suitable datasets for my applications
- ☐ I do not trust it
- ☐ Other, please specify

4. Are you a user of data arising from ground-based remote sensing, balloon-borne observations or aircraft observations? If so how do you use it, and if not why don't you? (please, choose all that apply)

a. Yes, I use these data for

- ☐ Model evaluation (all aspects including processes)
- ☐ Quality assessment of reanalysis data
- ☐ Comparison to other in-situ measurements
- ☐ Quality assessment of satellite raw data
- ☐ Quality assessment of satellite retrievals
- ☐ Assessment of long-term stability of satellite data
- ☐ Quality assessment of global/regional satellite-derived products
- ☐ Climate analysis (variability, trend detection, etc.)

- ☐ Other, please specify [ ]

b. No, I do not use these data because

- ☐ I do not need it
- ☐ I do not have access to it
- ☐ I cannot find it
- ☐ It is too complicated to use
- ☐ I cannot find the documentation
- ☐ The data format is not suitable
- ☐ I do not have suitable tools
- ☐ There are no suitable datasets for my applications
- ☐ I do not trust it
- ☐ Other, please specify [ ]

Uncertainties (3/5)

5. What is your level of expertise in using information about observational uncertainties related to satellite, ground-based, balloon-borne or aircraft data? (please, choose all that apply)

To activate the text field for the option "Other, please specify", please, choose the desired score first.

1 = little expertise, 5 = professional

	1	2	3	4	5
Gross uncertainties	○	○	○	○	○
Uncertainties arising from systematic effects of the measurements	○	○	○	○	○
Uncertainties arising from random effects of the measurements	○	○	○	○	○
Uncertainties arising from mismatch in measurement time, footprint and volume	○	○	○	○	○
Information about the representativeness of the measurement	○	○	○	○	○
Full uncertainty budget per observation system	○	○	○	○	○
Full traceable uncertainty estimates	○	○	○	○	○
Other, please specify [ ]	○	○	○	○	○

6. Do you use uncertainty estimates? (please, choose all that apply)

a. Yes

- ☐ Yes, I use uncertainty estimates

b. No, I do not use these data because

- ☐ I do not need it
- ☐ I do not have access to it
- ☐ I cannot find it
- ☐ It is too complicated to use
- ☐ I cannot find the documentation
- ☐ The data format is not suitable
- ☐ I do not have suitable tools
- ☐ I do not trust it
- ☐ Other, please specify

7. Which guidance on how to utilize observational uncertainty information related to satellite and sub- orbital data would be valuable to you? (please, choose all that apply)

To activate the text field for the option "Other, please specify", please, choose the desired score first.

1 = least preferred, 5 = most preferred

	1	2	3	4	5
Classical user guide document	○	○	○	○	○
List of peer reviewed papers	○	○	○	○	○
A help desk providing support on request	○	○	○	○	○
Online training course	○	○	○	○	○
Other, please specify	○	○	○	○	○

Data availability, format and tools (4/5)

8. What timeliness of the observational data do you need for your intended applications? (please, choose all that apply)

- ☐ Near real-time (3 hours)
- ☐ Delayed mode (1 month)
- ☐ Offline (1 year)
- ☐ It is of no importance
- ☐ Other, please specify

9. GAIA-CLIM is planning to distribute all its data in Netcdf4 format containing Climate and Forecast compliant metadata.

To activate the text field for the option "Other format preferences, please specify", please, choose the desired score first.

1 = least preferred, 5 = most preferred

	1	2	3	4	5
What is your preference for NetCDF4?	○	○	○	○	○
Other format preferences, please specify					
Format 1 [ ]	○	○	○	○	○
Format 2 [ ]	○	○	○	○	○
Format 3 [ ]	○	○	○	○	○
Format 4 [ ]	○	○	○	○	○
Format 5 [ ]	○	○	○	○	○

10. What kind of tools do you expect to be available on the Virtual Observatory? (please, choose all that apply)

To activate the text field for the option "Other, please specify", please, choose the desired score first.

1 = least preferred, 5 = most preferred

	1	2	3	4	5
a. Data search and discovery					
Geographical sub-setting	○	○	○	○	○
Temporal sub-setting	○	○	○	○	○
Variable sub-setting	○	○	○	○	○
Observing system type sub-setting	○	○	○	○	○
Surface type sub-setting	○	○	○	○	○
Other, please specify [ ]	○	○	○	○	○
b. Online data post-processing					
Averaging in space and/or time	○	○	○	○	○
Smoothing of vertical resolution	○	○	○	○	○
Regridding to a common resolution (vertical, horizontal, temporal)	○	○	○	○	○
Other, please specify [ ]	○	○	○	○	○
c. Online data analysis (statistical analysis, etc.)					
Comparison of collocated data sets at full and subset level	○	○	○	○	○
Anomaly calculations	○	○	○	○	○
Specific assessment of long-term time series, e.g., trends, stability tests	○	○	○	○	○

Other, please specify [ ] ○ ○ ○ ○ ○

d. Flexible forward (RT) modelling capability

Flexible forward (RT) modelling capability (e.g. to map from model variables to observed variables) ○ ○ ○ ○ ○

e. Data extraction

Collocated data files (full and subsets) ○ ○ ○ ○ ○
Comparison statistics ○ ○ ○ ○ ○
Other, please specify [ ] ○ ○ ○ ○ ○

f. Data visualization

Geographical maps ○ ○ ○ ○ ○
Individual or multi data set time series of data and analyses results ○ ○ ○ ○ ○
Plots displaying comparison statistics (scatter, box whisker, etc.) ○ ○ ○ ○ ○
Other, please specify [ ] ○ ○ ○ ○ ○

g. Other, please specify

Other, please specify [ ] ○ ○ ○ ○ ○

11. Which existing tools do you find useful to compare data sets? (please choose all that apply)

- ☐ NORS Portal
- ☐ NPROVS
- ☐ ICARE
- ☐ AEROCOM
- ☐ Other, please specify [ ]

GAIA-CLIM Workshops with external users (5/5)

Please, note that the remaining questions are optional to offer also anonymous submission.

12. Would you be interested in taking part in GAIA-CLIM workshops? If you wish to get notifications about the workshop by email, please, tick yes and enter your contact info in question 15 below. The first workshop will be held on October 6, 2015 in Rome, Italy.

- ☐ Yes
- ☐ No

13. Do you have any suggestions for topics to be addressed in more detail during the workshops within the GAIA-CLIM project?

14. Would you be interested in testing the tools developed in GAIA-CLIM? If you wish to be contacted concerning the testing, please, tick yes and enter your contact info in question 15 below.

- ☐ Yes
- ☐ No

15. If you wish you can leave your contact info here.

Name	
Last name	
Email	
Address	
ZIP code	
City	
Country	
Company / Organization	
Department	

Thank you for having participated in the survey!

Appendix B. Copy of email sent to CLIMLIST

=-=

CLIMLIST Mailing Number 15-4-44

Origin: Peter Thorne <peter@peter-thorne.net>

=-=

Dear CLIMLISTers,

as part of the Horizon 2020 program of the European Commission there is a new project GAIA-CLIM <http://www.gaia-clim.eu/> which is looking at improving our capability to use ground-based, balloon-borne and aircraft (collectively termed sub-orbital) measurements to characterize satellite measurements. Envisaged work includes efforts to map existing sub-orbital measurement capabilities, improve metrological understanding of these measurements, and better account for the effects of inevitable measurement mismatches between what the satellite and sub-orbital measurements actually measure.

From an end-user perspective the key outcome will be a web-based 'virtual observatory' facility providing the co-locations of satellite and high quality ('reference') sub-orbital measurements and their quantified uncertainties which shall be hosted by EUMETSAT.

We would like to ensure that we hear from as many potential users as possible early in the project as to what their likely needs are from such a facility in terms of its functionality, timeliness, data formats etc. etc.. This will ensure that we build the final interface so that the project outcomes are as useful as possible to the global research community as a whole.

If you are a potential user of such satellite - sub-orbital co-location data in current or possible future work and want to help shape our project's outcomes to meet your envisaged needs please go to <http://tinyurl.com/gaia-clim-survey/>. The survey consists of 15 questions and should take no longer than 15 minutes for you to complete.

Your participation and perspectives will be hugely valuable in informing the project and its outcomes. Although European funded the facility is intended to be a global resource and hence participation in the survey is welcomed from both within and outside Europe.

Many thanks

Yours

Peter Thorne (project coordinator) on behalf of all GAIA-CLIM participants