WORKING DRAFT

Input Data Requirements of EOS-Funded Interdisciplinary Science (IDS) Teams

Technical Paper—Not intended for formal review or government approval.

February 1996

Prepared Under Contract NAS 5-60000

RESPONSIBLE SCIENTIST/ENGINEER

02/20/96

Lori J. Tyahla, Science Specialist II, ESSi — Date
EOSDIS Core System Project

SUBMITTED BY

Dr. Joy Colucci, Manager, Science Office — Date
EOSDIS Core System Project

Hughes Applied Information Systems
Landover, Maryland

1.0 Introduction

1.1 Purpose

The Ad Hoc Working Group for Consumers was formed by members of the scientific community to address the specific needs of potential users of the ECS for the data planned to be archived within the system. Members from the scientific community include the co-chairs, Bill Emery (University of Colorado, EOS Data Panel) and G. David Emmitt (Simpson Weather Associates, EOS Data Panel), Bruce Barkstrom (CERES instrument team), William Rossow (IDS Investigator and EOSP instrument team), Steve Goodman (IDS Investigator and LIS instrument team), Sigfried Schubert (IDS Investigator), Dave Skole (IDS Investigator), and Gary Geller (ASTER instrument team). Others supporting the work of the AHWGC are: H. Ramapriyan, Gail McConaughy, Yun-Chi Lu (ESDIS), and Joy Colucci and Lori Tyahla (ECS Hughes Team).

The purpose of this document is to describe the results of the first major activity of the Ad Hoc Working Group for Consumers (AHWGC): to collect details regarding the data needs of the members of the 28 NASA-funded Interdisciplinary Science (IDS) Teams. This decision was based on the expectation that the IDS Teams will be the "first line" of users of the ECS and, as a group, will require a large volume of data from the system. In order to avoid confusion when describing the collection of data from the IDS Teams and the data products that the teams require, the word "information" will be used when referring to the information provided by the teams and the word "data" will be used when referring to the data products and parameters needed by the teams.

Each IDS Team was sent a customized packet of materials containing a summary of previously collected information and describing the type of information now sought by the AHWGC. Information sought by the AWHGC for each IDS Team was: parameters needed, desired spatial resolution and coverage, desired temporal resolution and coverage, preferred frequency and medium of data delivery, expected ordering mode, and whether the data was desired before the validation process had been completed. The customized packets of materials were distributed (one to each Principal Investigator of each team) in late June and early July of 1995.

The information received from the teams will be provided to the ECS developers to aid design in the following areas: data server design, distribution hardware sizing, system operations, and system performance (see sec 2.2.1 for details). This document describes the method of information collection, the analyses performed on the information, and the results obtained.

1.2 Organization

This document consists of three main sections and one appendix. Section 2 details the development of the method for gathering the necessary information from the IDS Teams and includes a list of all materials sent to the teams to fulfill the request. Section 3 presents each analysis performed on the information - both methods and results. Section 4 provides a summary of the results and the conclusions that can be drawn from them. Appendix A contains the information itself that was analyzed.

1.3 Review and Approval

This Technical Paper is an informal document approved at the ECS Office Manager level. It does not require formal Government review or approval; however, it is submitted with the intent that review and comments will be forthcoming.

The ideas expressed in this Technical Paper are valid for six months from the approval date. Questions concerning distribution or control of this document should be addressed to:

Data Management Office
The ECS Project Office
Hughes Applied Information Systems
1616 McCormick Dr.
Upper Marlboro, MD 20774

2.0 Development of Information Collection Method

2.1 Introduction

In order to minimize the impact on the schedules of the IDS Teams, all previously existing information regarding their data needs was assembled and analyzed for appropriateness to the task at hand. The Science Processing Support Office (SPSO) at Goddard Space Flight Center (GSFC) provided the information that it had collected from the teams at various intervals over the past few years. This list of IDS Team needs can be found in a document produced by the Science Office within the Earth Science Data and Information System (ESDIS) Project at GSFC in April 1995 (*Output Data Products, Processes and Input Requirements, Volume III, Version 3.0, Draft, Appendices N and O*). Appendix N is a list of data parameter needs for each IDS Team and Appendix O contains a mapping of the needs in Appendix N to the planned data products from the EOS platforms. Since the information in these Appendices had been collected over the past few years, it was decided that all of it should be updated for the task at hand. The information from the SPSO was used as the basis for the method of information collection.

2.2 Method of collection

2.2.1 Determination of Content of AHWGC Request

The SPSO information was provided to the ECS Hughes Team in soft copy form. Appendix O of the SPSO document was separated into 29 separate spreadsheets - one for each IDS Team. Recall that this Appendix provided the mapping of the Teams' needs to the planned data products from the EOS instruments. The information in Appendix N for each team was then added to the 29 separate spreadsheets for completeness. This process resulted in 29 customized spreadsheets containing the previously-specified data needs of each individual team.

The spreadsheets were then examined to determine if they contained columns to collect the information that the ECS developers needed. The original spreadsheets from the SPSO document contained the following 10 columns:

product name

product ID

parameter name

parameter ID

units

absolute and relative accuracy

temporal resolution

horizontal resolution and coverage

vertical resolution and coverage

comments

It was decided that some of this information was still required, while more detail was necessary in some cases. Several versions of the spreadsheets were developed and reviewed by the co-chairs of the AHWGC, a few additional members of the scientific community, and the ESDIS and ECS personnel. The final AHWGC spreadsheets contained the following 18 columns (explained below):

ECS

parameter name

product number

product level

priority

platform

available horizontal resolution and coverage

required horizontal resolution and coverage

available vertical resolution and coverage

required vertical resolution and coverage

available temporal resolution

available temporal coverage

required temporal resolution and coverage

required frequency of data delivery

required medium and mode of delivery

volume per delivery

expected ordering mode

receive before validation is complete?

The purpose of the column called "ECS" was to denote which products previously requested were products that the ECS was responsible for distributing. This column was used by both the teams and in the analyses. If an "X" appeared in the "ECS" column, it signaled to the teams that a description of the product was available in the materials sent with the packet (see section 2.2.2.2). It also meant that completing the columns for that product was mandatory for the teams. This is because the ECS developers are interested in sizing the ECS to distribute the products in the ECS archive only; it is expected that if a user requires data not in the ECS archive, the data center that holds that data (NOAA, for example) will distribute the data directly to the user.

The "Parameter Name" column was required to obtain information about the *parameters* of interest to the teams, as opposed to products of interest. In general, a product consists of more than one geophysical parameter (such as sea surface temperature, snow cover, etc.). Information at the parameter level is required to determine the extent to which parametric subsetting must be performed on a product. This is especially true in the case of products containing a large number of parameters, such as the CERES products, some of which contain over 50 parameters. If a user is interested in one parameter only, then the system must extract the desired parameter values from the product before sending to the user, thus reducing the amount of unnecessary data that the user receives.

The column "Product Number" refers to the product numbers assigned to the planned data products from the EOS instruments or numbers associated with the currently existing non-EOS products listed in the Science Data Plan (July, 1994). This column was provided to the teams to enable them to search the soft copy list of products included in their packets to determine if they still required a product they had previously requested. It is also used in the analyses to determine volume.

The "Product Level" and "Platform" columns were provided in response to members of the scientific community who stated that this information would aid them in determining which products would be most appropriate for their needs.

The "Priority" column was included on the sheets to separate the essential data needs of the teams from data needs that would enhance their research, but was not critical to it. Thus, there were two choices that the teams selected for this column, "critical" and "desired enhancement". The results presented in this document are separated into these two categories.

The "Available" and "Required" coverages and resolutions were requested for several reasons. The most critical need for this information was to determine how many of the parameters requested by the EOS-funded teams would require subsetting and/or subsampling. If a large amount of subsetting/subsampling is required, more processing power will be needed to perform these operations. The information was also needed to determine data volumes before and after subsetting/subsampling to estimate the impact of these services on the teams. In other words, what reduction in data volume delivered to the user (significant in some cases) can be achieved by providing these services? It was also suggested that this information could be used to determine if the resolutions and coverages of the planned EOS products meet the requirements of the EOS-funded teams.

The columns regarding frequency of delivery, mode and medium of delivery, volume per delivery, and expected ordering mode were not requested in the past, but the information is needed by the ECS developers. The area most affected by this information is the sizing of the distribution hardware and the number of operations personnel required to place the required data onto the required media.

The AHWGC also requested that the teams denote whether they are interested in receiving each data product before the validation process is complete. The availability of the product in time is affected by the answer to this question. If the team prefers to wait until the validation process is complete, they cannot begin to receive the product at the time it is initially available to users.

Table 2.1 is a summary of the information requested of the IDS Teams and the areas of the system design that are impacted by the team responses.

Table 2.1 Design impacts of team responses to the AHWGC request for information

Information provided by IDS Teams	Result of analysis	Design Impact
Priority; Temporal Coverage and Resolution	Realistic relative pull on data products for each of three time periods	Data Server Design - disk space, cpus, read/write heads; Archive Structure - on-line vs. near-line; Processing Requirements - cpus, standard production vs. on-demand
Horizontal and Vertical Coverages and Resolutions	Subsetting requirements	Data Server Design - cpus, read/write heads, server workspace System Performance - response times
Frequency of data delivery and Volume per delivery	Data volume delivered vs. time	Data Server design - staging area requirements
Medium and Mode of delivery	Number and types of media; preferred methods of delivery	Distribution Hardware Sizing - tape/media drives, read/write heads, pieces and types of media, ftp sites; System Operations - number of personnel
Expected ordering Mode (ad hoc vs. subscription, etc.)	Number of concurrent users vs. time of day	Data Server Design - staging area requirements, cpus System Operations - distribution of data during off- peak hours System Performance - response times

Upon final approval by the AHWGC of the information content of the spreadsheets, the next task was to determine for which time frame(s) the IDS Teams should identify their data needs. A yearly breakdown was suggested, but the final decision was to break the time period into increments that are tied to releases of the ECS and launches of the EOS platforms. At the same time, the desire was to limit the number of time intervals to a number that was easily manageable by the teams. The resulting time frames and corresponding milestones are listed below:

Period 1: December 1, 1996 to August 30, 1997 (Begins with Release A of ECS and includes the nine months leading up to the launch of the TRMM platform),

Period 2: September 1, 1997 to June 30, 1998 (Begins with Release B and the launch of TRMM - very close in time - and includes the ten months leading up to the launch of the AM-1 platform),

Period 3: July 1, 1998 to June 30, 1999 (Begins with launch of AM-1 and includes the twelve months following the launch).

In order to facilitate ease of completion of the response for the teams, each team's customized spreadsheet was separated into three spreadsheets - one for each time frame. For example, if a product from the original customized sheet was not available until the third time period, it was removed from the sheets for the first two time periods. This reduced the occurrence of teams requesting data products before they were available.

2.2.2 Preparation of Packet of Materials

Upon completion of each team's set of customized spreadsheets, a packet of supporting materials was prepared. The packet included both hard- and soft-copy versions of the set of customized spreadsheets (as described above), a catalog of data product descriptions, a soft-copy of the April, 1995 version of the list of planned EOS parameters, a soft-copy of the tables from the July, 1994 version of the Science Data Plan (Schwaller and Krupp, *eds.*), instructions for computing data volumes and a sample calculation, sample areas (in km^2) of representative portions of the Earth, and a list of terms and codes used in the spreadsheets and their meanings. Two letters were also included - one from Ghassem Asrar (EOS Program Manager, NASA Headquarters) and one from the co-chairs of the AHWGC, explaining the purpose of the request. The details of each item are explained below.

2.2.2.1 Cover Letters

Two cover letters were sent with each packet. The first was written by Ghassem Asrar on behalf of the Science Data Panel endorsing the activity of the AHWGC and requesting that the teams review their data needs and respond to the AHWGC request. This letter also encouraged the teams to provide "believable and defensible" estimates of their needs. The second letter was written by Bill Emery and Dave Emmitt (co-chairs of the AHWGC) and Lori Tyahla (ECS User Characterization Team Lead). This letter explained the need for the information and included Table 2.1 of this document.

2.2.2.2 Catalog of Data Product Descriptions

During the initial review of the spreadsheets, several investigators indicated that data product descriptions would be necessary for them to choose appropriate products. In April, 1995, the ECS User Characterization Team placed a product use survey on a data server and made it accessible to potential users via the World Wide Web (WWW) (see doc # 161-TP-001-001 for details of survey). This EOSDIS Product Use Survey contained descriptions of the planned EOS data products as well as data products that were expected to migrate from the Version 0 system during Release A of the ECS. These descriptions were pulled together and arranged by instrument and placed in a separate document titled, *ECS Catalog*. The descriptions in this document reflected the best information available as of May 18, 1995.

2.2.2.3 List of Planned Data Products from EOS Instruments

The SPSO at GSFC compiled a spreadsheet list of planned parameters to be produced from the EOS instruments (April, 1995). This table is Appendix G in the document titled, *Output Data Products, Processes and Input Requirements, Volume II, Version 3.0, Draft*. In the spreadsheet list, each parameter is mapped to a specific data product number; the same product numbers are used in the *ECS Catalog*. Thus, team members were able to search the soft-copy parameter sheet for parameters of interest, note the product number the parameter is contained within, then refer to the catalog of descriptions for more information about that product.

2.2.2.4 Tables of Data Products from Science Data Plan

The Science Data Plan (July, 1994) contains tables specific to each of the currently existing data centers that will become Distributed Active Archive Centers (DAACs) in the EOS era. The individual DAAC tables were extracted and merged into one large table. A soft-copy of this merged table was placed on the ECS Science Office homepage (http://ECSInfo.hitc.com) and instructions on how to access it via the WWW were included in the packet. This spreadsheet table enabled the teams to search for currently existing data that they might continue to require. Supplying their needs for this data was optional for two main reasons. First, it was not clear at the time of the distribution of the packets when each data set would be migrated from the existing data center into the ECS archive. Second, it was decided by the members of the AHWGC that requesting the teams to supply information for these products as well as the planned EOS products would cause too much of an impact to the research schedules of the teams.

2.2.2.5 Instructions for Completing Spreadsheets

A one-page set of detailed instructions for providing the requested information was provided to each team in order to obtain the results in a consistent manner. The teams were asked to specify the frequency at which they require their data to be delivered and to compute the volume per delivery that they expect to receive. The method given to them to compute volume was meant to be simple and generic. The teams were to determine the number of square kilometers in their area of interest, determine the number of "resolution cells" that would be required to cover the area, based on the available resolution of each parameter of interest. The next step was to assume that each resolution cell would contain two bytes per parameter. Then, based on the required temporal coverage and the available temporal resolution, the team calculated the expected volume per year for each parameter. The final step was to divide the total volume per parameter per year by the number of deliveries per year to obtain a volume per delivery for each parameter. A list of sample areas (for example, the area poleward of 60 degrees latitude = 3.42×10^{7} km^{2} per pole) was provided. The sample areas were chosen based on the areas of interest identified by the teams in the EOS Reference Handbook (1995).

The last instructional item in the packet was a list of definitions of each column heading in the spreadsheets and for any and all codes to be used in the spreadsheets (for example, the teams were to enter "DE" in the "Priority" column for a parameter that was a desired enhancement).

2.2.3 Distribution of Packets

All packets were prepared and distributed by the ECS Hughes Team. The soft-copy version of the set of customized spreadsheets was placed on two floppy disks - one formatted for use with a PC, the other formatted for use with a Macintosh computer. Separate arrangements for soft-copy delivery were made for teams that could not use either of these disks. The packets were sent via Federal Express to each team's Principal Investigator and should have been received by mid-July, 1995. Each team was to coordinate and provide one response to include the data needs of all members of the team.

3.0 Analysis and Results

3.1 Introduction

As of February 20, 1996, responses to the AHWGC request were received from 13 of the 28 teams. For the most part, the teams followed the instructions for providing the requested information. However, some inconsistencies in the responses across the teams were noted, and requiring further preparation before the analysis could proceed. Examples of irregularities are: one team provided the words "<full data set>" for the volume per delivery instead of the actual volume; one team listed 1 GB as the volume per delivery for each parameter they expect to receive; some teams specified more than one type of preferred media for a parameter, some teams did not use the proper codes for priority, some teams specified entire products instead of parameters, etc. The total number of products/parameters identified is 819.

Most of the inconsistencies were easily remedied; however, one unexpected problem arose during the analysis of the data that was not solved as quickly. When examining the expected volume of data per delivery for one of the teams, the numbers appeared to be much smaller than the Hughes Team would have expected. The volumes for this team were recalculated based on the parameters and spatial and temporal coverages and resolutions specified by the team. It became clear that the generic method of calculating volume (see section 2.2.2.5) underestimated the actual volume; the underestimate was quite significant for several of the parameters. Overall, this team was found to have underestimated the volume of data by approximately one order of magnitude. However, the team in question has not yet been notified of the recalculated (larger) volumes and, once apprised of the situation, may revise their data needs to reduce the volume of data to a more manageable amount. It is possible that the volumes for other teams may also be underestimated due to the method of calculation.

The set of 13 team responses was merged into three files - one for each of the three time periods. Due to the impending Critical Design Review for Release B (scheduled for April, 1996), the analysis for time period 3 was completed first. To date, the following results have been produced from the team responses: subsetting/subsampling needs, data delivery frequency, media preference, ordering mode, and total volume to be delivered to all 28 teams. The following sections describe the methods employed and present the results of each type of analysis.

3.2 Subsetting and Subsampling Needs

A product or parameter was noted as requiring spatial subsetting if the required spatial coverage (in km^2) was smaller than the coverage listed as available for that product or parameter. Similarly, a product/parameter will require spatial subsampling if the desired resolution was coarser than the available one. The third analysis category is whether or not the team required a resolution which is finer than the one currently planned.

The analysis proceeded as follows. Three columns were added to the merged spreadsheet for time period 3. If a product/parameter required spatial subsetting, a "1" was placed in the spatial subsetting column; if not required, a "0" was placed in this column. The same is true for the other two categories. An Excel macro was then written to sum up the number of products/parameters that required each type of subsetting/subsampling. The results are shown in Figure 3-1.

Figure 3-1 Subsetting/subsampling needs of 13 responding IDS Teams

It appears that, overall, the planned spatial resolutions and coverages will meet the needs of these 13 teams. The "Unknown" category appears for products and parameters where no desired resolution or coverage was provided by a team; therefore, it is not known if the planned resolutions and coverages meet the needs of the teams. One exception to this is the Mark Abbott team. The team did not provide requirements for resolutions but asked that we assume that the available ones meet their needs. The distribution in Figure 3.1 indicates a more widespread need for spatial subsampling than had previously been thought. In addition, for a small number of cases, the planned resolutions of some of the products/parameters are not fine enough for the needs of the teams.

3.3 Delivery Frequency

The teams were asked to specify one of the following delivery frequencies for each product/parameter of interest: daily, weekly, monthly, quarterly, annually, or other. Some teams provided a number of deliveries per year, such as 48. All delivery frequencies were converted to number of deliveries per year according to Table 3.1.

Table 3.1 Number of deliveries per year for delivery categories

Delivery category	Number of deliveries per year used for yearly volume calculation	Range of number of deliveries per year used for distribution by category
Daily	365	> 100
Weekly	52	26 to 100
Monthly	12	7 to 25
Quarterly	4	2 to 6
Annually	1	1

Converting the delivery categories to number of deliveries per year was necessary to compute the total volume of each parameter to be delivered per year. However, in order to produce a distribution of delivery frequency preferences by the above 5 categories, all frequencies specified by the teams as a number of deliveries per year were converted to one of the categories. These conversion factors are also listed in Table 3.1. An Excel macro was written that binned the requests according to the third column in Table 3.1; the results are shown in Figure 3-2.

Figure 3-2. Delivery frequencies based on number of requested products/parameters.

In this distribution, the category "other" includes "unknown" (not specified by team) and "as available" (send as soon as it's been produced). It is interesting to note that over 80% of the data requested by the 13 responding teams is expected to be received on a monthly basis. This large number of monthly deliveries will impact operations planning and staffing at the DAACs. Several planning options exist for fulfilling these requests. The first option is to fill the orders at the end of each month, placing a heavy load on the data staging areas as well as on the DAAC staff. The second option is to fulfill a specified number of monthly orders per day, thus distributing the workload over the entire month. Undoubtedly, other solutions are also possible; it will be the task of the Maintenance & Operations (M&O) segment to propose plans for this activity and to implement one or a combination of solutions.

3.4 Media Preference

The IDS Teams were asked to specify one of the following media types for each requested product/parameter: tape, CD-ROM, electronic, or other. The AHWGC determined that asking the teams to specify particular types of tape (4 mm, 8 mm, etc.) would impact the research schedules of the teams, so the more general category of "tape" was employed. The category of "electronic" is equivalent to "ftp". The word "electronic" was used because it is a medium, whereas "ftp" was seen as a "mode" of delivery; in the resulting distribution, "ftp" is used as a medium. An Excel macro was written that examined the appropriate column in the spreadsheet and summed up the results for each category as shown in Figure 3-3.

Figure 3-3. Media preferences based on number of products/parameters requested by 13 responding teams.

The category "unknown" in Figure 3-3 includes cases where no medium was specified. This category is shown in the distribution for completeness. Generally speaking, there appear to be no major surprises in this distribution. If one adds the "tape" and "CD-ROM" percentages, one arrives at the conclusion that 62% of the requested data (by number of parameters and NOT by volume) will be required on media and about 37% will be received over the network (1% is unknown). One can perform an analysis on this data to product a distribution based on the volume of data to be received on media vs. over the network; this type of analysis is not included in this document at this time but is planned for the final draft.

3.5 Ordering Mode

There are three main methods by which users can order data from the ECS: interactively, via subscription, and by automated processes. Interactive ordering implies that the user is connected to the ECS and through the user interface, provides the information for the specific data he or she desires.

A subscription order is one that has been set up by a user during a previous interactive session where he or she has specified the data required as well as the frequency at which he or she would like to receive it, such as weekly or monthly. It is expected that the user will receive an e-mail message notifying him or her that the data is available on at a particular ftp site, and also providing the user with a password for access to it. The user then accesses the site and "pulls" the data to his or her machine.

Ordering via an automated process occurs when a piece of software, or other non-human entity, places orders for data on behalf of a registered user. This enables researchers to place individual data orders at a speed unachievable by a human user. This functionality is most probably a Release C capability; however, it is included here for completeness.

The distribution for ordering mode for the 13 responding IDS Teams is shown in Figure 3-4.

Figure 3-4. Ordering mode based on number of requested products/parameters

In Figure 3-4, the category "unknown" applies to cases where no ordering mode was specified by the team. Again, this distribution is what one would expect from the IDS teams. As the acronym "IDS" implies, the data received by the teams is spread across many disciplines and will be used to study global change. Many of the teams are developing models and will need regular deliveries of data to provide input to these models, so it is not surprising that more than 50% of the data products/parameters requested will be needed at regular intervals by the teams. About 20% of the data will be ordered interactively, and about 11% will be ordered via an automated process.

3.6 Yearly Volume of Data Required by IDS Teams

As stated before, the IDS teams are expected to pull a large amount of data from the ECS. The actual volume pulled has, until now, remained unknown. There are two major obstacles in computing the yearly volume to be delivered to the teams. The first is the underestimate of volumes due to the imperfect method of volume calculation (see section 3.1), and the second is the fact that only 13 of the 28 IDS teams (or 46%) provided input to this analysis. The volumes presented in this analysis are based on the *available* data product resolutions; this leads to an overall overestimate of the total volume to be delivered because about 30% of the products/parameters will be subsampled (reducing the volume to be sent) before delivery. The final draft of this document will contain a detailed analysis of the impacts of these two offsetting volume estimation errors.

The AHWGC decided to extrapolate the information provided by 13 teams to 28 teams for an initial analysis. The basis for this extrapolation is presented in the sections following the analysis of the 13 responses.

3.6.1 Analysis of Inputs from 13 Responding Teams

Strictly speaking, each row in the merged spreadsheet for all teams was to contain the details for one individual parameter; however, not all teams specified their needs at the parameter level. For those cases that are specified by parameter, some teams chose to include the total volume in the volume per delivery column and placed "0" for the volume of other parameters from the same product. For example, the Rothrock team requires 8 parameters from the CER14 product. The volume per delivery in the row containing the first parameter is actually the total volume for all 8 parameters; thus the volume per delivery in the rows for the other parameters in this product is equal to 0. The volume for any parameters containing "0" for volume per delivery is assumed to be included in the first non-zero volume preceding the zero volume.

An Excel macro was written to compute the total volume expected to be pulled by each team on a yearly basis. The volume per delivery was multiplied by the number of deliveries per year; then these volumes were summed over the parameters for each team. The resulting yearly volume of data distributed to each team is shown in Figure 3-5, arranged from smallest volume to largest. Figures 3-6a and 3-6b show the yearly volumes for critical data and desired data, respectively.

Figure 3-5. Total yearly volume of data required by 13 responding teams (includes both critical and desired data).

Figure 3-6. Yearly volume of data required by 13 responding teams for a) critical data needs and b) desired data needs.

3.6.2 Extrapolation of Volumes to 28 Teams

Extrapolating the volumes shown in Figures 3-5 and 3-6 to include the needs of all 28 teams is only valid if the 13 responding teams adequately represent the entire group. Three parameters were examined to determine if this was the case: volume per team, research disciplines and geographic scale of team, and number of team members.

3.6.2.1 Volume per Team

Figures 3-5 and 3-6 indicate that, with respect to required data volumes, the group of 13 teams span a large range of require data volumes, the smallest total yearly volume (Figure 3-5) is 1.01

GB/year for the Sellers team and the largest total yearly volume is 23,786 GB/year for the Lau team (remember that these are the volumes *before* subsampling is performed).

Figures 3-6a and 3-6b indicate that, for some teams (Grose, Wielicki and Lau), all of the data is critical. For the Tapley team, the volume of the desired data greatly exceeds the volume of the critical data. For most of the other teams, the volume of critical data is greater than or equal to the volume of desired data. Thus, based on the volume of data requested, the distribution of the 13 responding teams can be extrapolated to all 28 teams.

3.6.2.2

Table 3.2 shows disciplines of interest, geographic scales of research, and number of team members for the 13 responding teams. The source of this information is the 1995 EOS Reference Handbook. Table 3.2 displays a wide range of all three parameters and supports the conclusion that the data volumes for the 13 responding teams can be extrapolated to represent the group of 28.

Table 3.2 Disciplines of interest, geographic scales of research, and number of team members for the 13 responding teams

Team	Number of Members	Disciplines	Scale of Effort
Hartmann	6	Air/Sea Interactions	Local, Regional, Global
Isacks	7	Land/Climate Interactions, Hydrology, Geomorphic Processes	Local
Tapley	10	Angular Momentum Budget - Coupling between air/sea/land	Global
Rothrock	11	Interactions of Ocean, Ice, and Atmosphere	Regional
Sellers	11	Biosphere/ Atmosphere Interactions	Global
Abbott	11	Atmosphere/Ocean Interactions and Ocean Primary Production	Regional, Global
Grose	13	Atmosphere - Radiation, Chemistry, Dynamics	Global
Dickinson	16	Hydrology, Radiation, aerosols, atmosphere	Local, Regional, Global
Mouginis-Mark	16	Volcanology, Atmospheric Chemistry	Local, Regional, Global
Barron	18	Hydrology	Local, Regional, Global
Wielicki	19	Radiative Energy Balance	Local, Regional, Global
Godfrey	24	Ocean, Biological Processes	Regional, Global

3.6.3 Yearly Volumes of Data Distributed to all 28 Teams

If one sums the data from Figures 3-5 and 3-6 across the 13 responding teams, the resulting volumes are as follows:

Total data volume (13 teams)	=	66.2	Terabytes/year
Critical data volume	=	40.7	Terabytes/year
Desired data volume	=	25.5	Terabytes/year

In order to extrapolate the volumes to represent all 28 teams, each volume was multiplied by the factor (28/13), the ratio of the total number of teams to the number of responding teams. The following volumes result:

Total data volume (28 teams)	=	142.6	Terabytes/year
Critical data volume	=	87.7	Terabytes/year
Desired data volume	=	54.9	Terabytes/year

3.7 Data Products Requested by 13 Responding IDS Teams

The 13 responding IDS teams have requested a total of 140 distinct products. In addition, 25 requests were unspecified as to which product should be the source of the desired parameter. Table 3.3 contains the number of requests (at the *product* level) for each of the products requested by the 13 responding teams. Of the 25 "unknown" product requests, 14 are truly unknown, 2 were associated with ASTER products, 9 with CERES products, and 3 with products from the Data Assimilation System (DAS), referred to as the "DAO" in the table. Products that do not have the "standard" EOS product number (such as CER02, AST04, etc.) can be found in the Science Data Plans according to the product numbers in Table 3.2. The one exception to this is the group of products from the TRMM platform - the data product numbers shown in Table 3.2 correspond to those in the ECS Catalog (see section 2.2.2.2).

One must remember that the products requested in Table 3.3 reflect the needs of 13 of the 28 teams. Although the *volume* of data requested by the 13 responding teams can be extrapolated to represent all 28 teams, the individual data products requested by the teams **can not.** Thus, the number of requests for each product in Table 3.3 **can not** and **should not** be used to alter the list of planned products.

Table 3.3 Number of requests by 13 responding IDS teams for each of 140 distinct products.

Product #	Number of Request	Product #	Number of Request	Product #	Number of Request	Product #	Number of Request	Product #	Number of Requests
ASTER	2	DAO	3	SAG02	1	TRMM Products		J-32	1
AST02	2	DFA02	1	SAG03	1			J-33	1
AST03	4	DFA03	1	SAG05	1	TM1-1	2	J-34	1
AST04	3	DFA04	1	SAG06	1	TM1-2	1	J-40	1
AST05	3			SAG07	1	TMICP-1	1	J-41	1
AST06	1	MIS01	1	SAG08	1	TMICPV-2	1	J-42	1
AST07	3	MIS02	1	SAG09	1			J-49	1
AST08	4	MIS03	3	SAG10	1	VIRS-1	2		
AST09	5	MIS04	6						
AST10	1	MIS05	7	SWS01	1	PR-2	1	L-49	1
AST13	2	MIS08	3	SWS02	1	PR-5	1		
AST14	3	MIS09	1	SWS03	4	PR-7	1	LST701	2
		MIS10	1						
CERES	9	MOD01	1	MOD20	2	A-4	1	M-13	1
CER01	2	MOD02	5	MOD21	2	A-5	1	M-15	1
CER02	1	MOD03	3	MOD22	4	A-12	1	M-44	1
CER03	3	MOD04	8	MOD24	2	A-15	1	M-45	1
CER04	4	MOD05	7	MOD25	1	A-26	1	M-64	1
CER05	3	MOD06	11	MOD26	2	A-27	1	M-73	1
CER06	4	MOD06A	1	MOD27	2	A-28	1	M-85	1
CER07	3	MOD07	3	MOD28	7	A-32	1	M-89	1
CER08	3	MOD08	3	MOD29	3	A-33	1	M-91	1
CER09	1	MOD09	6	MOD30	8	A-34	1	M-108	1
CER11	4	MOD10	6	MOD31	2			M-109	1
CER12	1	MOD10B	1	MOD32	2	E-2C	1	M-110	1
CER13	2	MOD11	9	MOD33	3	E-5	1	M-111	1
CER14	2	MOD12	5	MOD34	1			M-112	1
CER15	2	MOD13	3	MOD35	3	G-1	1	NCD-71	1
CER16	1	MOD14	2	MOD36	1	G-5	1		
		MOD15	4	MOD37	2			SI-6	1
		MOD16	1	MOD38	2				
		MOD18	2	MOD41	1				
		MOD19	2	MOD42	2				
Unknown	14								

4.0 Summary and Conclusions

4.1 Summary and Conclusions

Although responses were not received from all 28 IDS teams, the AHWGC decided to extrapolate the 13 responses to represent all 28 teams for an initial analysis. The total yearly volume that the 28 teams will receive is on the order of 145 Terabytes/year where about 90 Terabytes/year (or about 62%) of the total volume is deemed to be critical data for the teams.

Spatial subsetting and subsampling of data prior to delivery will be of great use to the IDS Teams. About 30% of the products/parameters requested will require subsampling (14% unknown) and 13% will require spatial subsetting (14% unknown). Parametric subsetting is also required by the teams. Although a detailed analysis is not presented here, it is obvious when examining the requested data in the merged spreadsheet that not all parameters within a particular product are requested.

Monthly delivery of data is, by far, the most popular frequency at which to receive data (65% of the number of products/parameters requested). Delivering data at this frequency will be convenient for the teams, but will require ECS developers to plan for an efficient method of handling this workload.

More than half (62%) of the number of products/parameters are desired to be delivered on some type of physical medium with the remaining products/parameters to be delivered over the network. Tape is about three times as popular as CD-ROM for the IDS teams.

Slightly more than half of the requested products/parameters (54%) are requested to be received at regular intervals via a subscription service. This is a significant portion of the number of requested products/parameters and indicates a clear need for the ECS to provide this type of service. A small, but significant portion (11%) of the number of requested products/parameters will be ordered via an automated process. This results points to the need for the ECS developers to continue to consider implementing this feature.

A wide variety of products have been requested by the 13 responding IDS Teams. The total number of distinct products identified by the 13 responding IDS Teams is 140; 25 products/parameters requests had no product information. The range of products requested is large and reflects the interdisciplinary nature of the investigations. However, one **can not** and **should not** use the information in Table 3.3 to alter the list of planned products because the individual data products needs of the 15 non-responding IDS teams can not be inferred from this table.