Using online data for independent research

How to succeed at university in GEES disciplines

Using online data for independent research

Richard Waller, School of Physical and Geographical Sciences, Keele University
David M. Schultz, School of Earth, Atmospheric and Environmental Sciences, University of Manchester

Contents

1. Introduction

1.1. Context and problem

Independent research is an important and distinctive component of degree programmes within the GEES disciplines, most commonly encountered within final-year projects and dissertations. Successfully completing a large research project with relatively little support requires strong self-motivation, effective time management, problem-solving skills, writing and presentation skills, and a high degree of organisation. As such, it can be one of the most challenging and daunting elements of your degree programme.

The ever-growing range of digital information, that is freely available at the click of button, is a fantastic resource that can be used to enhance your project. Online datasets and images can be used in conjunction with traditional field and laboratory-based methodologies. Alternatively, your project may revolve exclusively around the location, manipulation and analysis of secondary data. The successful use of online data for independent research is however associated with a number of challenges.
This online resource has been designed to help you to identify and deal with many of the problems commonly encountered when using online data within the context of independent research. More specifically, the following sections will help you to:

- locate relevant data;
- determine whether it is fit for purpose and available for you to use;
- present your data clearly and effectively.

1.2. Online data and information literacy

The skills and approaches central to the location and effective use of online data can be considered as a subset of the skills known more broadly as *information literacy*. Information literacy involves the creation, gathering, usage, management, synthesis and creation of information and data in an ethical manner and can be broken down into the following stages (modified from SCONUL, 2011):

1. Identification of a need for information.
2. Assessment of current knowledge and the identification of gaps.
3. **Construction of strategies for locating information and data.**
4. **Location and accessing of the required information and data.**
5. **Comparison and evaluation of the data collected.**
6. **Organisation of information professionally and ethically.**
7. **Application of the knowledge gained: presenting results, synthesising data to create new knowledge and disseminating in a variety of ways.**

The stages of the process that are emboldened are of particular relevance to independent project work and the use of online data. They can be used as a guide and a reminder of the issues to bear in mind when using any kind of secondary data. Their importance will be explored further within this resource.

1.3. Value of these skills to the GEES graduate and future employers

When undertaking a prolonged research project on a specific topic, it is easy to become lost in the detail and to forget about the whole host of skills you are acquiring at the same time. Reflecting on these skills and how they might be of value to you personally, and to future employers, provides an excellent way of providing the motivation required to complete a major research project. The skills most commonly developed include:

- **the ability to locate, utilise, synthesise, manipulate and present relevant information and/or data** (the focus of this resource) – this is likely to incorporate the effective use of information technology and the development of numerical and graphical skills;
- **competency in the use of specific techniques** – these techniques range from the appropriate use of particular laboratory or field equipment to the development of questionnaire surveys;
- **problem solving skills –** in particular, the application of knowledge and skills to "real world" problems;
- **production of a professional report** (and possibly an oral presentation of your results) – communication skills.

Thinking beyond the confines of academic curricula, it is striking to note how closely the broader generic skills associated with independent work map onto the competencies most highly sought after by employers (Brennan *et al*., 2001):

- working under pressure;
- accuracy and attention to detail;
- time management;
- adaptability;
- initiative;
- working independently;
- taking responsibility and decisions;
- planning, co-ordinating and organising.

Consequently, it is clear that independent project work enables you to acquire and develop a wealth of abilities that will be of value to a range of employers. It is therefore well worth reflecting upon and taking careful note of the various skills you are developing while working on your project. This information can then be used to good effect within your personal CV and application forms that can otherwise lack the concrete evidence on your abilities that employers are often searching for. Being able to talk clearly and enthusiastically about your independent project work is also likely to impress once you have made it to an interview.

Once again, considering the potency of your independent project work as a means of integrating and showcasing the talents that you have developed during degree programme will provide a powerful immediate source of motivation for its successful completion and your subsequent move into graduate employment.

2. How to find data sources that you need for your research project

The most immediate problem encountered when trying to use secondary data within an independent project is locating an appropriate source of the data for your chosen task. This section focuses on the issues commonly encountered during this initial stage of your research, the questions you need to ask yourself, and the ways in which typical problems can be overcome. Search strategies that can be used to locate appropriate sources for your data are addressed in section 2.1 (section 4 provides a list of potential data sources of particular relevance to the GEES disciplines that can be used as a starting point or a source of inspiration). Once you have located a promising dataset, sections 2.2 and 2.3 cover the issues you need to consider to determine whether it is fit for purpose and permissible for you to use. Finally, section 2.4 highlights the need to acknowledge your sources and how best to reference any datasets you have employed.

2.1. Locating sources of geophysical and geographical datasets

With more digital information and data being freely available than ever before – literally at the touch of a button – you might think that acquiring the information you need should be straightforward. However, with such a profusion of material available online, locating the *specific* dataset that is right for *your* purpose can be really challenging. It is very easy to waste hours of your time in what can seem to be an increasingly fruitless search. Therefore, to save time and frustration, it is important that you carefully consider and devise an appropriate search strategy. In this respect, searching for a dataset is, in essence, no different to finding that elusive journal article for an essay assignment, and the techniques you can use are very similar. Some of the strategies you can use are listed below:

Find any relevant organisations: identifying organisations that collect and/or archive the kind of information you need provides an ideal place to start your search (see section 4 for some inspiration). For example, if you need gauging data or water quality information on a particular river, then the Environment Agency might be an obvious place to look. Once you have found an organisation, have a trawl through their webpages to see if they have an obvious site which provides access to information or even data downloads. Some national organisations have webpages specifically designed for students wanting to use their data in independent projects (http://www.metoffice.gov.uk/education/teachers/past-weather-data). In other situations, finding the right webpage can be much harder. If you have no joy, then you can always try to contact an appropriate person at the organisation to ask for help and advice. If you do contact someone, ensure that you provide some background information on your project and be as specific as possible about the data you require. (An email phrased along the lines, "please send me all your information on…" is unlikely to elicit a response!) Conversely, saying that you are working on a project with "Dr. Supervisor" and promising to send them a copy of your final project shows you are happy to provide something in return and may increase your chances of getting a positive response. It could even result in a job opportunity further down the line.

Using data portals: research and data collection activities that are supported by public funds are increasingly being made publically available as a condition of the funding. Therefore there are an increasing number of portals that have been set up by research organisations and other bodies that provide access to a range of datasets. The British Geological Survey provides an excellent example of this approach that can make your life a lot easier: http://www.bgs.ac.uk/opengeoscience/. Further examples of these data portals are provided in section 4.

Ask your supervisor: as part of your project, you will have been allocated a supervisor who may well be an active researcher in the subject area you are addressing. Therefore, if you are struggling to find any relevant sources, it is well worth seeking their advice. However, it is important to show that you have done some legwork yourself first. In particular, if they run a module on the same topic, check your notes and other materials just in case they have already highlighted some sources you might use.

Contact a particular researcher in the field: as part of a broader search of relevant literature, you might identify a particular author who has collected data that appears to be relevant to your project. You could try to contact the researcher directly, in which case you should again be polite and specific in your request. Do not be surprised if you receive no response, but you might just get lucky.

General Internet search: If these more targeted approaches don not work or are not appropriate, then simply searching the Internet is your final option. In order to have a reasonable chance of success, then you need to put plenty of thought into the search terms you are going to use. Key things to consider are:

1. **Persevere:** it is unlikely that you will find a whole range of relevant data sets within five or ten minutes of starting your search. However, if you persevere and spend an hour or two experimenting with different search terms, then you should start to identify some useful leads. In other words, you need to <u>invest some time</u> to get a worthwhile end result. Do not wait until it is too late to start your project.

2. **Be specific:** if your search term is too general then you will generate more hits than you can possibly hope to sift through. A quick Google search for "climate change data", for example, yields 395 million results. Therefore, while this approach might yield some of the major sources or suppliers of data, obtaining specific data for a more specific project needs a more careful approach.

3. **Keywords:** the secret to finding relevant data is the thoughtful and creative use of various "keywords" as your main search terms along with what are known "Boolean operators" (e.g. AND, OR etc.). For example, if you were searching for data on the changing incidence of extreme weather in the UK, then you could include these terms within your search term:
   "climate change data" **AND** "extreme weather" **AND** "UK" **OR** "United Kingdom"

Providing a specific location or agency likely to have the data may help narrow your search. If you are looking for data from specific instrumentation, list that in the keyword search.

2.2. Determining if your data is "fit for purpose"

Once you have identified a potentially useful online data source, the next stage is to check whether it is fit for your intended purpose. The first thing to consider, as with any source you employ, is whether the data has come from what you consider to be a reliable and trustworthy source. In most cases, the data should be attributable to an organisation or individual that has a reputation within the field. If a dataset has appeared as part of a more general Internet search and you are unable to identify the source, then you need to be more cautious.

Once you are happy that the data is reliable, the key question is whether the dataset has the requisite scope for your project. In this respect, scope might refer to the spatial extent of the data collected, although more commonly it will relate to its temporal characteristics, notably its length and its resolution. Focusing on the latter, if you were aiming to examine the impact of post-war catchment changes on the discharge of a river system and the dataset only contains data for the last ten years, it clearly does not have the *length* of record required. Equally, if you were aiming to look at the changing incidence of flash flood events over a period of time and the data you acquired only contained monthly averages, then it would lack sufficient resolution.

If at this stage you have a dataset that you do think is fit for purpose, then section 3.1 will provide a range of useful tips on how to work with your data. If you conclude that your data is not fit for purpose, then you can always contact the organisation or individual concerned to check whether any additional data is available. In this respect it is important to note that organisations will often only provide a certain amount of data for free. Therefore you might need to carefully balance your requirements for data with your willingness to incur additional charges.

2.3. Making appropriate and ethical use of your data

An important but frequently neglected aspect of using online data relates to issues surrounding copyright protection, licensing and intellectual property rights. Once you have identified what you think is a worthwhile data set, it is very important that you carefully check the terms and conditions associated with its use. This will enable you to use the data legally and ethically and may involve you having to complete a declaration and return this to the supplier before the dataset is released. Common requirements relating to the use of the data involve agreements to use the data only for the stated purpose; not circulate the data to other parties; and to delete the data after a given period of time.

On a positive note, there has been an increasing move towards promoting fuller and freer access to data. Rather than the standard "all rights reserved" copyright statement, organisations in North America in particular are allowing access through "creative commons" licences (http://creativecommons.org/about). In addition, a number of organisations in the U.K. have simplified their arrangements for using online data by allowing access through an open government licence (http://www.nationalarchives.gov.uk/doc/open-government-licence/). This removes the requirement to submit an application in return to agreeing to abide by the licence agreement. Unfortunately, the terms and conditions associated with the use of online data are not always as clear and transparent as you would expect. If you are uncertain, you should contact the supplier directly.

One thing to bear in mind when looking into licensing agreements is that they are generally much more open to educational and "not for profit" uses. In the commercial sector, you may well find that numerous sources available to students are no longer free.

2.4. Acknowledging and referencing your sources

An almost universal requirement regarding the use of data is that you acknowledge its source within any related publications. If you are using a dataset obtained online, you are likely to be required to adopt a particular attribution statement in any results that include its use (e.g. "Contains British Geological Survey materials ©NERC [year]").

If you are using more tangible products such as maps or images, then these should be referenced in the normal way, just as you would with any other information resource such as a book or journal article. To reference an electronic image obtained from the Internet using the Harvard style of referencing would involve:

Author, Year (image created). *Title of work*. [type of medium] Available at: include URL [Date accessed]

3. Effective use of data

Starting a research project, especially with a large dataset with many different variables that *could* be analysed, can be intimidating. Yet, there are approaches that can be useful for getting started and maintaining focus, which will be discussed in section 3.1. Tips on how to use statistics properly are given in section 3.2. Finally, section 3.3 offers guidance about how to effectively present your graphics.

3.1. Getting started

Reduce the size of the dataset to only those that are most important or most relevant: some variables may not be relevant to the problem you are investigating. If so, then you may wish to create a reduced dataset that contains just the relevant variables. Doing so will reduce the size of the dataset, occupy less space on your computer, allow you to see the relationships among the elements of the dataset more clearly, and help you maintain focus on your specific research question.

Reduce the size of the dataset by eliminating incomplete records: for many datasets, all the essential variables that we might want to analyse may not be included. For example, a dataset of tornadoes across Europe may lack the precise time that the tornado was observed. If the purpose of the study was to investigate what time of the day tornadoes are observed, then tornado records missing the reporting time cannot be used. That does not mean that those reports are not valuable for other parts of the study (say, the most common months for tornadoes). Therefore, make sure to retain those records.

Become comfortable with making plots of what the data look like: plot a histogram of the distribution of the quantities so that you can see the mean, minimum, and maximum of the fields. Plot the points on a map so that you can see the spatial distribution of the data. Although you may not use these plots in your report, do not underestimate the importance of getting familiar with your data. Avoid the rush to immediately start creating plots for the report without doing this step.

Check for unrealistic values in the dataset: is the relative humidity greater than 100%? Are there data points in the middle of the ocean? Are concentrations of chemical species less than zero? If so, then you may want to examine why those data points exist. Were the raw data entered into the dataset incorrectly? Was the instrument malfunctioning? Were the raw data fed into an algorithm that was not properly tested?

When dealing with large datasets, **it often helps to break the datasets into subsets:** and compare data from within those subsets. For example, if you have a list of urban floods from the past ten years across the globe, floods in New York City may occur under different circumstances than floods in Delhi. Many Earth processes may occur by more than one mechanism, and understanding the variability is an important part of understanding the dataset as a whole. For example, although all the volcanoes on Earth have some general similarities, the closer you examine them, the more the differences become apparent. Volcanoes occur in different geological environments – island arcs associated with subduction zones such as the Aleutian Islands of Alaska, hot spots such as the Hawaiian Islands, and continental supervolcanoes like Yellowstone. The geochemical data and emitted gases vary in those different environments. Even within a class of volcanoes, there will be variability. Thus, before synthesizing your results across the whole dataset, it may be worth looking at subsets of the data first and looking for groupings that allow you to categorize them.

It may be useful to just take one or two events and perform an in-depth case study first: the depth that you can place on one case may help you see the important processes at work. Then, having a clear picture of how one case works, you can try to extend that to more elements of your dataset. It may be easier to generalize your results to the larger dataset after having looked at a smaller number of cases in more detail.

Creating plots of the data is just the first step in doing research: it is an important step, but creating a plot must be followed up by analysis of the plot, interpretation of the plot, relationship to any previous studies that have created similar plots, and then implications of those results for your study as a whole. If you have the data that can answer your questions, then do not speculate wildly. Use the data in your study to test your speculation.

3.2. Pitfalls to avoid with statistics

You may have heard Mark Twain's quote: "There are three kinds of lies: lies, damned lies, and statistics." Although statistics can be used to cleverly provide authority to situations where the evidence is not particularly strong, statistics also provide a powerful tool to demonstrate legitimately the strength of your research results.

If you have ever picked up a statistics book, the large number of statistical tests and formulas may be quite daunting. Indeed, many smart scientists can be humbled by the field of statistics. The goal of this section is not to provide a complete guide to using statistics in your research, but to give some advice about the common pitfalls and misconceptions that you can easily avoid. Correct application of statistics in your report can be the difference between a 2.2 and a 2.1, or a 2.1 and a First.

3.2.1. Linear regression

Many spreadsheet applications such as Excel come with an easy way to assess whether the data fits a line, a process called *linear regression*. Linear regression is used to look for relationships between variables and assess the strength of that relationship. But, linear regression can be misused in several ways. Specifically, students often fit their data to this line, even if their data is not best explained by that linear fit. Let us explore this point in more detail.

1. **Linear regression is not useful for all datasets:** consider the data in Figure 1. The data within each of the four panels all have the same best-fit linear regression line ($y = 3x + 0.5$) and the same correlation coefficient ($r = 0.82$). Despite their quantitative similarity, panels (b), (c), and (d) are inappropriate uses of linear regression. In panel (b), the data fits a perfect parabolic trajectory described by a quadratic equation, not a linear equation. In panel (c), the data would perfectly fit a different line if the outlier at $x = 13$ were excluded. In panel (d), the correlation in the dataset is entirely due to the one outlier at $x = 19$. Only in panel (a) is linear regression an appropriate statistical tool.

Anscombe's Quartet

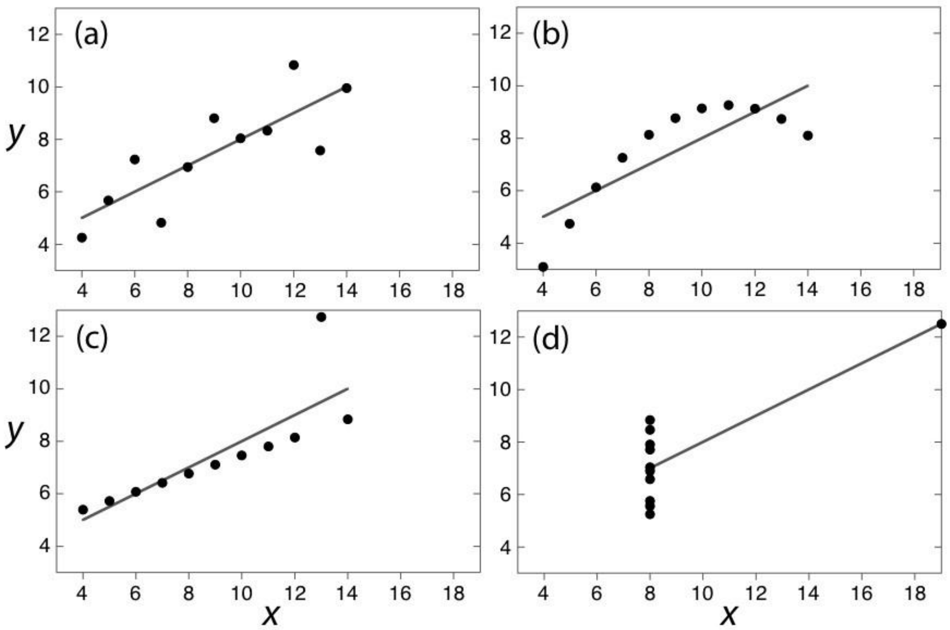


Figure 1: Anscombe's (1973) quartet – four examples of datasets that have the same mean (7.5), same standard deviation (4.1), same linear regression line (y = 3x + 0.5), and same correlation coefficient (0.82). Only (a) is an appropriate use of linear regression. (Figure and caption from Schultz (2009, Figure 11.12).)

2. **The correlation coefficient is a measure of the strength and direction of the relationship between *x* and y:** if all the data were perfectly correlated along a line, then $r = 1.0$. A perfect negative correlation (*y* decreases as *x* increases) would be $r = -1.0$. Uncorrelated data have $r = 0$. Strong correlation coefficients are typically greater than 0.8 (or less than –0.8, for negative correlations), and weak positive correlation coefficients are typically less than 0.5. Thus, just because you can fit a line to the data does not imply a strong physical relationship between the two variables.

3. **The square of the correlation coefficient tells you something about the strength of your relationship:** the square of the correlation coefficient (called the *coefficient of determination* r^2) assesses the percentage of variance explained by the linear fit. That means for the data in Figure 1a, $r^2 = 0.6724$, or 67% of the variability in the data is explained by the linear fit. Knowing the value of r^2 tells the reader the proportion of the variability in *y* that is predictable from knowing *x*.

4. **Correlation does not equal causation.** Consider the example in Figure 1a with $r^2 = 0.6724$. Of course, this correlation does not mean that *X* caused *Y*, but it is suggestive of a relationship between the two. All you can say is that the two have a mathematical similarity. It may be that the two variables are linked, but not to each other. For example, the number of tornadoes reported over the United States has been increasing since the 1950s (Verbout *et al*., 2006). In addition, the carbon dioxide in the atmosphere has also been increasing (http://keelingcurve.ucsd.edu). Yet, the number of tornadoes is not related to the increase in carbon dioxide. Tornado reports are increasing due to better observing systems such as radars and greater emphasis on data collection (not necessarily more tornadoes), and carbon dioxide is increasing largely because of burning fossil fuels.

3.2.2. Dealing with noisy data

Many geophysical data can be noisy. Data with high temporal frequency or high spatial density can often indicate real variability in the data that is simply too small to measure adequately. Alternatively, instrument error can also be quite large, depending on the type of instrument and the specific measurements taken.

There are many different ways to smooth or filter the data; some are better than others. The so-called box-car smoother (also called running-mean smoothers) average 3, 5, 7, 9, or more neighbouring points to reduce the amplitude of noisy peaks. Unfortunately, such smoothers may take outlier data and spread its effect to surrounding points. Such smoothing may produce false frequencies in the data that are not physically meaningful (called *aliasing*). If the outliers can be identified and eliminated, then smoothing has a better chance of producing meaningful values. Identifying outliers is another reason why taking a hard look at the properties and quality of the dataset as a whole before jumping in too deeply into analysis is a good idea (section 3.1).

An example of how failing to remove outliers before smoothing results in making the outlier problem worse is demonstrated in Figure 2. Not removing the outlier before applying a three-point smoothing function results in the two surrounding points being abnormally increased. Removing the outlier results in a much more reasonable result.

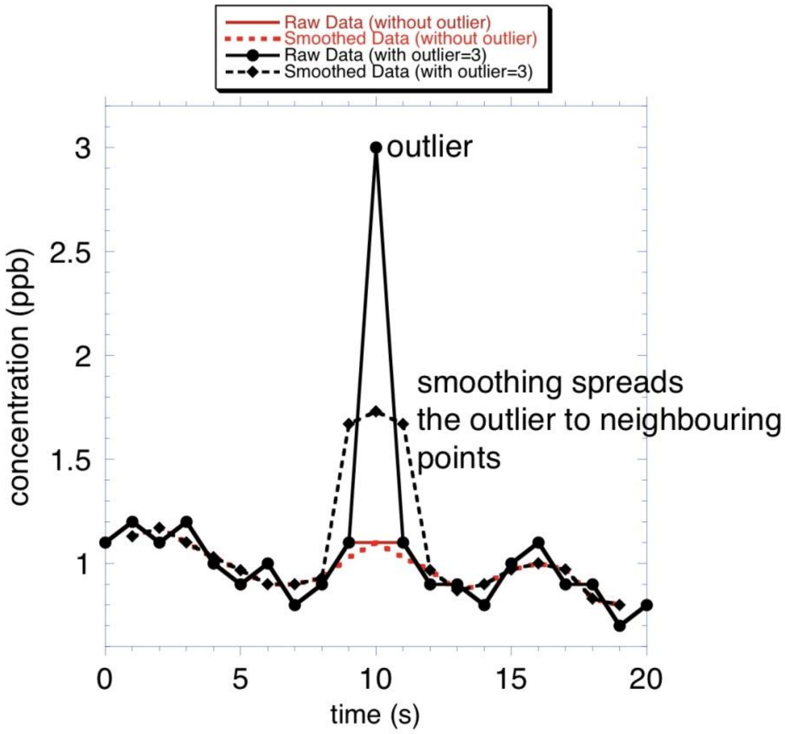


Figure 2: Data from an instrument measuring concentration of a chemical constituent once a second. The raw data (solid black line) contains an outlier at 10 s. Applying a 3-point smoother (equal weighting among the three points) aliases the outlier to the two surrounding data points (dashed black line). Removing the outlier (solid red line) before smoothing (dashed red line) alleviates the aliasing.

Another issue is how to deal with calculations involving noisy data. Two noisy datasets when multiplied or divided by each other can produce extreme values that are not physically meaningful. In this way, small amounts of variability in the data can amplify to yield lots of variability. Clearly, you want to avoid making the problem worse, so in most cases filter or smooth noisy data first before performing calculations that will amplify the variability. A superior alternative would be to determine the slope of the relationship from a linear regression.

3.2.3. A special note about Microsoft Excel

Microsoft Excel has enabled just about anyone to organize data, perform calculations, and create data plots. As powerful as Excel is, it can also be limiting or even incorrect. The default plots that it creates may need to be improved (section 3.3). The option to smooth lines sometimes produces unusual results. Some calculations that it performs may even be incorrect. In fact, in 2008, a whole section of the peer-reviewed scientific journal *Computational Statistics and Data Analysis* discussed the errors in the numerical algorithms that perform calculations within Excel (McCullough, 2008). The problems with Excel notwithstanding, use the plots it creates carefully and avoid default plots, if possible.

3.2.4. Using mathematical terms correctly

There are a number of mathematical and statistical terms that are often used improperly. This section presents a collection of some of the more common misuses.

Correlate/correlation: be careful when using these words in the sense of “relate” or “relationship”. Readers may infer that a correlation coefficient has been calculated. Use “relate” or “relationship” instead.

Resolution versus grid spacing: when the resolution of a dataset is described, oftentimes the writer is referring to the length of the grid intervals between data points. For example, a regularly spaced Cartesian gridded dataset representing a digital elevation model (often abbreviated to DEM) may be represented by 1km square grid boxes. Strictly speaking, the resolution of this dataset is actually 3–8 times the grid spacing of many datasets. That is because it would take 3–8 grid points to identify (or *resolve*) the structure of a feature. For example, Figure 3 shows that five grid points at 5-km intervals are needed to begin to approximate the shape of the sine wave with 20-km wavelength. Therefore, use the term *grid spacing*, unless you really do intend to mean the scale of the features that can be resolved.

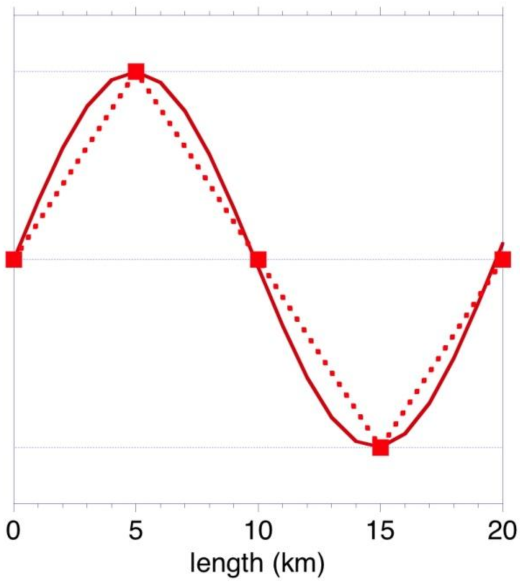


Figure 3: A figure demonstrating that five grid points are needed to begin to resolve a sinusoidal feature with a wavelength of 20 km. (Figure adapted from http://www.meted.ucar.edu/mesoprim/bandedprecip/print.htm.)

Significance/significant: be careful when using these words in the sense of "substantial" or "important". You do not want readers to infer statistical significance when none has been calculated. It is best to avoid this term.

Standard deviation versus standard error: do not confuse the terms *standard deviation* and *standard error*. *Standard deviation* is the measure of the spread of the data away from the mean of a sample and is a measure of the variability of the dataset. The standard deviation is relatively insensitive to the sample size. On the other hand, *standard error* is for expressing an uncertainty in a reported statistic, such as a mean. For example, the standard error can represent the standard deviation of the sample mean and is a measure of how accurate the estimate of the sample mean is from the population mean. As sample size increases, the standard error decreases.

***t* test:** this is formally known as the "Student *t* test", not the "student *t* test". Student was the penname of author William Sealy Gosset, a chemist working at the Guinness factory, who developed the test in 1908 as a way to determine the quality of stout from a sample population. The *t* test is used to quantify the likelihood of whether two sample datasets could be derived from the same overall population of data.

3.3. Effective presentation of data

When constructing plots of our data, we often rush through to see what the data looks like when finally plotted. That is, after all, the excitement of discovery that we feel when doing science. These initial plots may explain to *us* what is going on, but they often need some editing and revising before they can explain what is going on to the *reader* (in your case, someone who might also be the *marker*). For example, Figure 4a is what comes out of a software plotting package when all the default values are set. However, the meaning of this figure is not clear. What the values of *X* and *Y* represent is not known. The lines of data are almost indistinguishable from the default grid lines of the figure. This figure is our *working figure* or *draft figure*. It is useful for our purposes as we understand the data, but it should not be submitted in your report, as the reader would be lost.

In contrast, Figure 4b is what is called a *publication-quality figure*. Creating this figure required about 20 minutes of additional effort to make the individual data lines distinguishable and labelled directly. In addition, the axes are labelled and now are meaningful to the reader, tick marks are added, and the grid lines are removed.

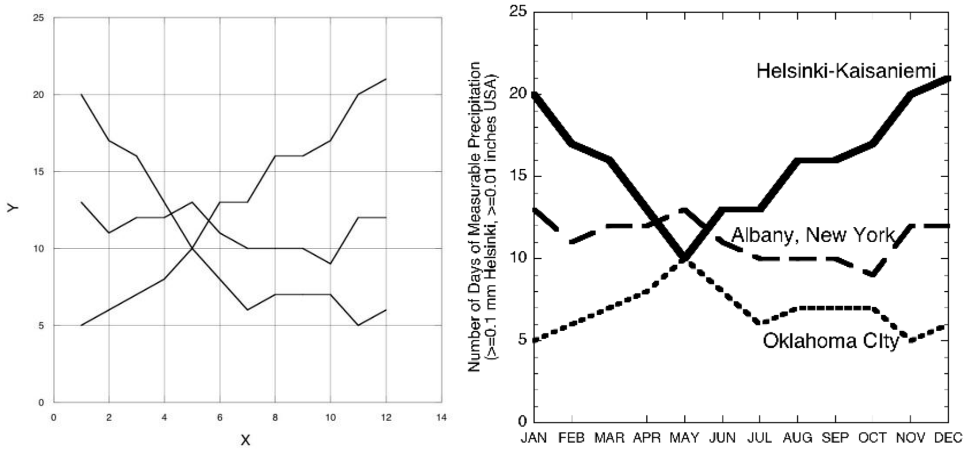


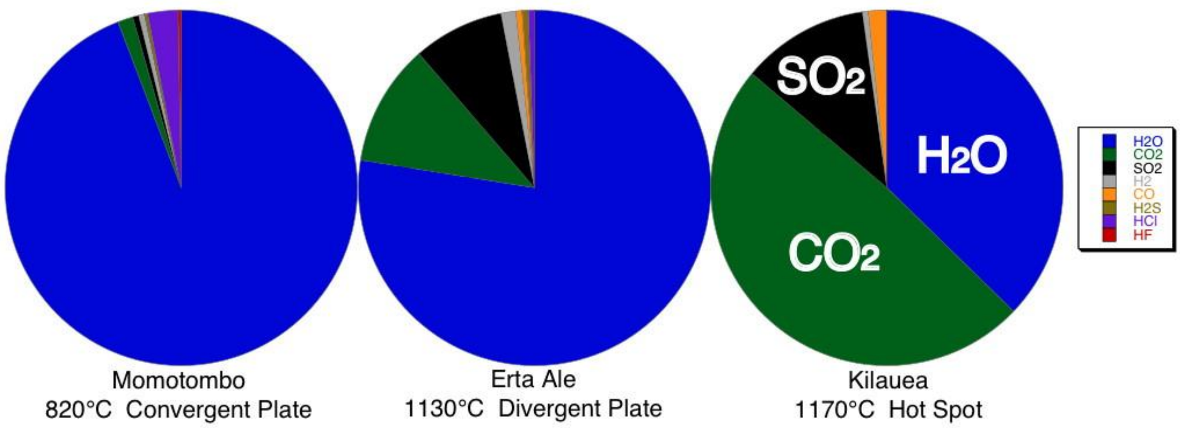
Figure 4: (a) Draft figure, (b) publication-quality figure. (Figures from Schultz (2009, Figures 11.5 and 11.6).)

When you are preparing your report for submission, allot yourself some time to create the figures. Carefully creating figures that clearly represent the data demonstrates professionalism in your work.

3.3.1. Constructing convincing figures

To convince a reader of the correctness of your argument, you have both the text of your report and the accompanying plots of the data in the form of figures. Relying on the default figures that your software generates may not be the best way to present your data. Here are several examples demonstrating that by choosing a different figure or designing the figure differently you can construct a more convincing figure. And, a more convincing figure is more likely to receive a higher mark.

For our first example, consider the three different measurements of gases from active volcanoes in Figure 5. The data are presented two separate ways: three pie charts (Figure 5a) and a stacked bar chart (Figure 5b). Although both are acceptable means of presenting the data, presenting the data as three stacked bars is a more compact representation. Moreover, the bar charts are easier to obtain quantitative information from.



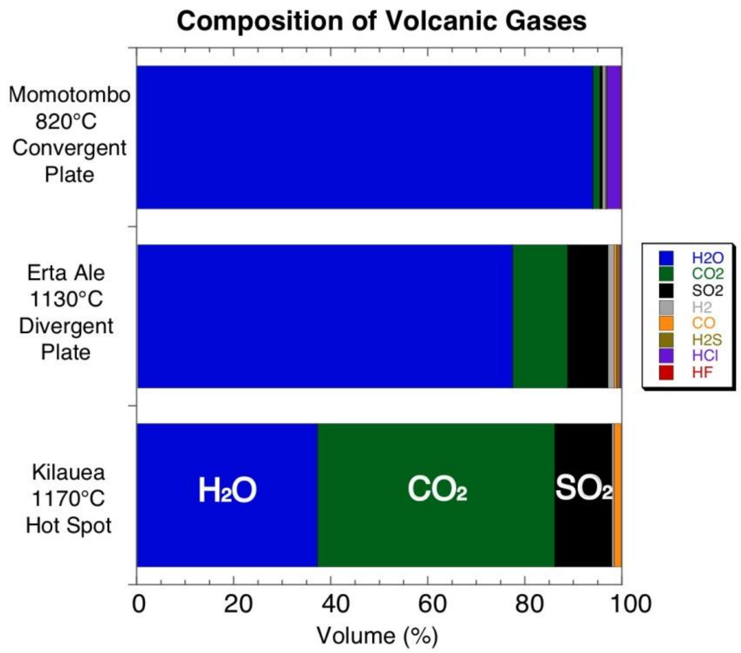


Figure 5: Comparing (a) pie charts to (b) bar charts. White labels for the gases have been annotated to assist readers in recognizing the dominant gases without having to read the legend. (Data taken from http://volcanoes.usgs.gov/hazards/gas, originally from Symonds et al. 1994.)

For a second example, consider two different ways to present the data in Figure 6. On the left, the bars are ranked in alphabetical order: cell phone to WeatherBug (Figure 6a). On the right, the bars are ranked from the black bar with the largest value to the black bars with the smallest value (Figure 6b). Which is easier to read and interpret? Clearly it is panel (b). Because the black bars are ranked from largest to smallest, we might expect that the grey bars would also rank from largest to smallest, as well. Such a ranking is largely apparent, except for two important differences. Both the grey bars for The Weather Channel and National Oceanic and Atmospheric Administration (NOAA) Weather Radio do not follow the trend of decreasing percentage from top to bottom, indicating that these two sources were sometimes preferred over other comparably ranked sources of weather information. Figure 6 demonstrates that a little extra effort in re-arranging the data before constructing the figure can be quite helpful to the readers and may even expose hitherto unknown information about your dataset.

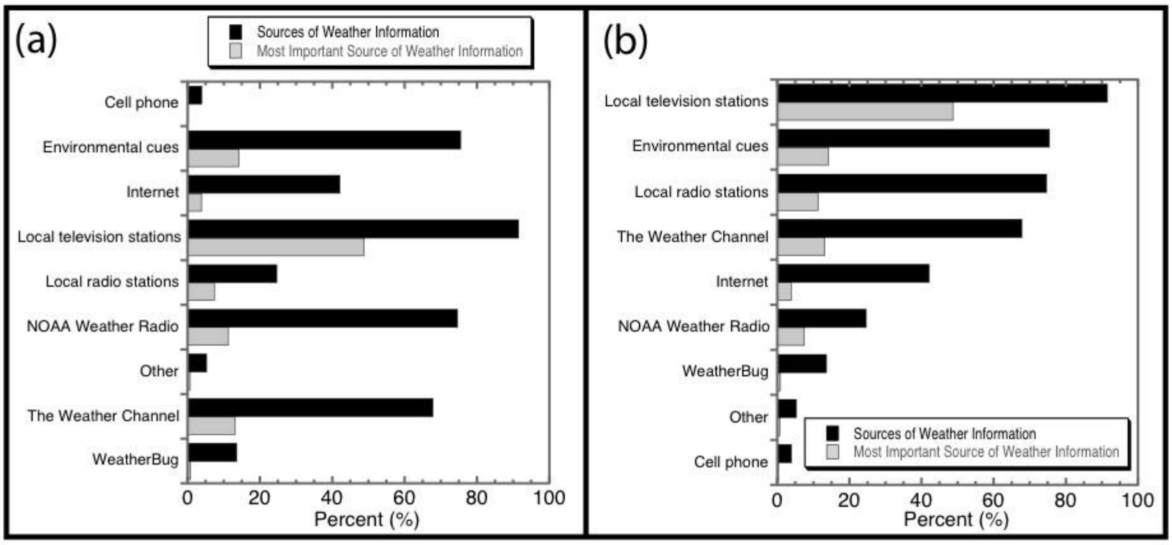


Figure 6: The importance of thoughtful ordering of bars to produce an effective bar chart. (a) alphabetical ordering of sources leads to an ineffective bar chart. (b) ordering of sources by decreasing percentage produces a more effective bar chart. (Figure and caption from Schultz (2009, Figure 11.14).)

It is often said that if you cannot see the difference between the distributions visually, then the results are often not statistically significant. A third example demonstrates the utility of *box-and-whisker plots* to compare three different distributions of data and get a feel for the similarities or differences in different groups of data without formally computing tests of statistical significance. In Figure 7, measurements of sand-body dimensions in fluvial data from three different layers are plotted. The box surrounds the middle 50% of the data (also called the *interquartile range*), within which are the median (black vertical line) and mean (grey vertical line). The whiskers of each box extend away from the data at a value of 1.5 times the interquartile range. The dots below each box-and-whisker plot show the distribution of the data; these dots are not always shown in these kinds of figures. Guidance on how to create box-and-whisker plots can be found in Banacos (2011).

The implication from Figure 7 is quite clear: the Molina member has the largest width sand bodies of all the three members. These bodies are *substantially* larger in the Molina, with the lower end of the Molina member interquartile range exceeding the upper end of the interquartile ranges for the Atwell Gulch and Shire members. Such a graph can provide confidence that your results are statistically significant, even if the statistical test is not formally calculated.

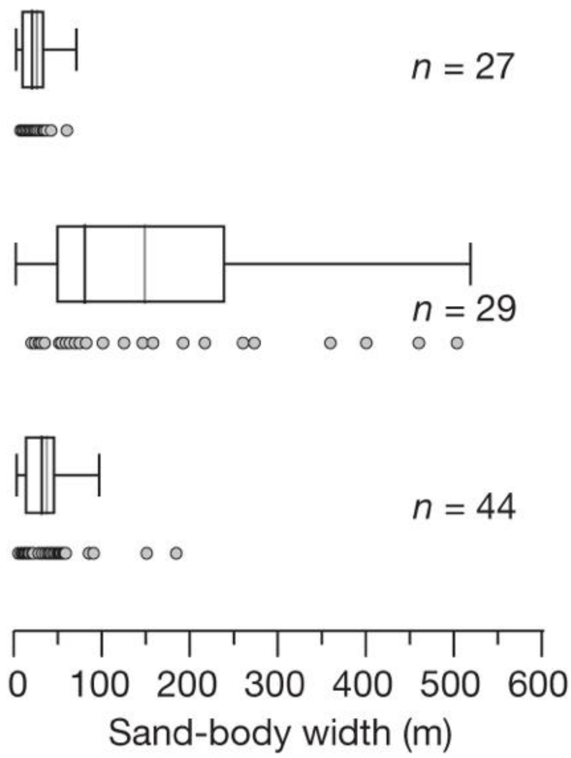


Figure 7: Box and whisker plots of fluvial data from the Atwell Gulch (bottom), Molina (middle) and Shire (top) members. Edges of boxes denote bounding quartiles, the black vertical lines represent median values, the grey vertical lines represent mean values, and whiskers denote the lower fence and upper fence (that is, 1.5 times the interquartile range). Grey circles denote individual data points. (Figure and caption adapted from Foreman et al. (2012, Figure 3b).)

Finally, we present a fourth example of how to construct convincing figures. After your research is completed, you may wish to **construct a schematic diagram** of your research to provide an overview of your project and to distil the essence of the work into a conceptual model. For example, geologists typically construct geological cross sections of their field area to synthesize and extend the structure seen on the surface below ground. A further step could involve drawing the structure of your field area at different times in the geological past to show its evolution. In another example, if you had described how the loss of Arctic sea ice has led to marked changes in the atmospheric flow patterns across the

Northern Hemisphere, you might draw a map of the regions affected, showing a schematic location of the anomalous jet stream (Figure 8).

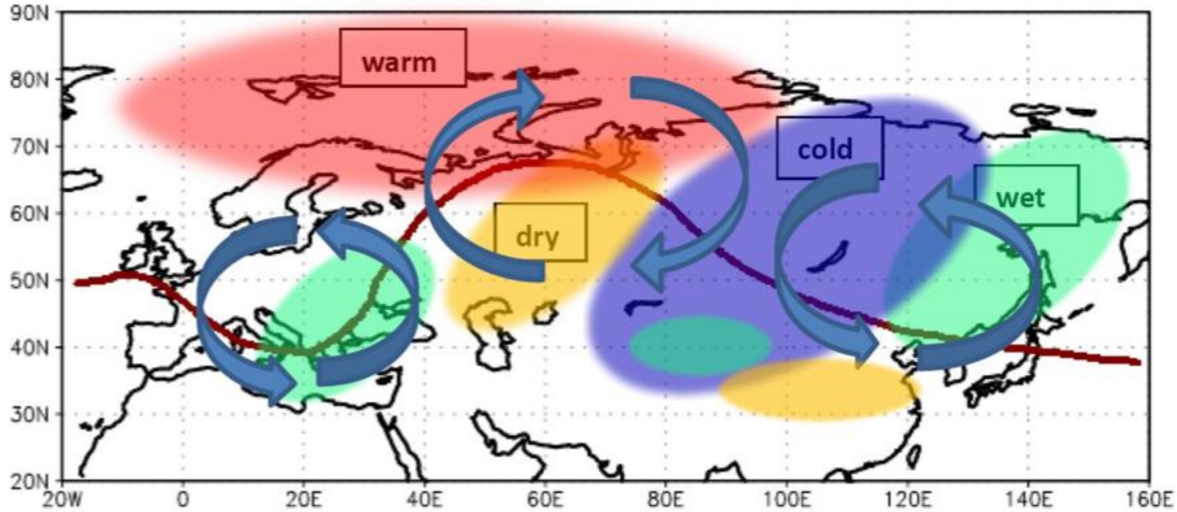


Figure 8: Schematic of how reduced Arctic sea ice affects winter surface air temperature and precipitation tendencies across Eurasia. Arrows denote the spatial distribution of an anomalous anticyclone and cyclones associated with the negative phase of the tripole wind pattern in the lower-troposphere, the brown line represents a 500-hPa height isoline, yellow and green areas indicate less and more precipitation, respectively, and red and purple areas depict positive and negative surface air temperature anomalies, respectively. (Figure and caption from Wu et al. (2013, Figure 13).)

3.3.2. Tips for producing effective graphs

When creating a plot for your report, ensure that your graph has its essential components, many of which are illustrated in Figure 9. Not all graphs will have these components, but many will.

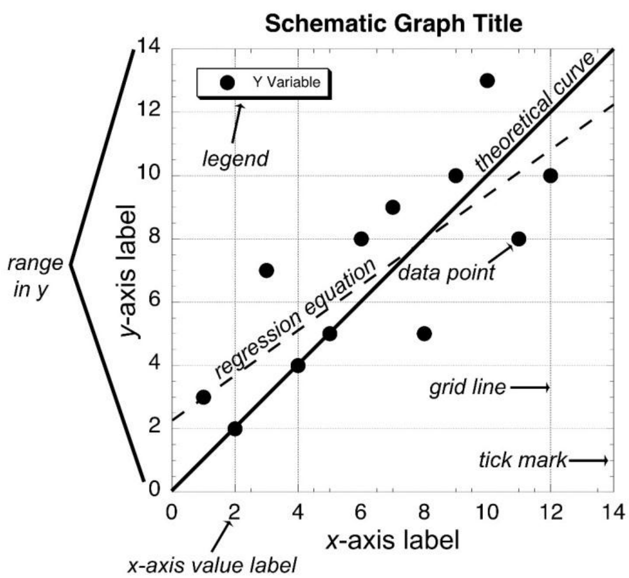


Figure 9: Schematic scatterplot with components labelled.

Here are some tips for constructing an effective graph:

1. **Datapoints:** ensure that the data points are large enough to be reproduced if the graph is reduced to fit into the report. Small dots may work for situations when hundreds or thousands of data points populate the graph, but make the dots larger when fewer data points exist.
2. **Axes:** do not be afraid to use nonlinear scales on axes (such as a logarithmic scale) if the data is more appropriately represented on such a graph.
3. **Labels on *x* and y axes:** ensure that all your axes are labelled, and the units are clearly stated. Use complete words so that the reader will understand what is being plotted, rather than the default headers of the variables in your spreadsheet.
4. **Axis values labels:** make sure these numbers are large enough to read when the figure is reduced to fit into your report. Again, make sure that the labels are meaningful. Rather than "*T* [°C]", write "2-m Temperature [°C]". Rather than "Grain size 1", write out the values of each category (e.g., "0.10–0.19"). Rather than numbering the months "1…12", label them "January…December".
5. **Ranges of axis values:** to reduce the amount of empty white space in your graph, consider reducing the range as much as possible while still encompassing all the data.
6. **Legend:** if there are multiple symbols or line types representing the data, consider including a legend. If the data are distinct enough, skip the legend and label the data points or lines separately, as was done in Figure 4b.
7. **Grid lines:** grid lines are often default in many plots generated from Excel. For scientific purposes, these grid points clutter the figure and can often be removed. However, specific grid lines annotated on the graph can be helpful to identify important values (e.g., "0°C").
8. **Theoretical curves:** if both the x and y data are related through some equation or theoretical curve, consider including this.
9. **Regression line:** more on regression lines is discussed in section 3.3.
10. **Shape of the figure:** if you are comparing two datasets that have equal ranges (such as comparing observed data to modelled data), make the graph square and include a 1:1 line to represent the ideal fit for the data.

Here is a checklist to make sure that you have constructed an effective figure:

1. **Does the data stand out from the figure?** Are the data points or data lines easily distinguished from the grid lines?
2. **Have you annotated the graph to ensure readability?** For example, for geologic maps, have you labelled specific features or locations you describe in the text so that readers are properly oriented (remember that not every reader will be familiar with your field area). Do you have a horizontal length scale, vector length scale, legend, colour bar, reference lines, title of the figure, error bars, confidence intervals, or labels for features described in text?
3. **Do the quantities plotted within the figures include units?**
4. **Are all figures in your report constructed consistently, where possible?** Are multiple-part figures all constructed with the same ranges, where possible? Are all base maps using the same domain? Are contour intervals consistent?
5. **Have you used scientific date/time format: "1200 UTC 10 April 2012"?** Formats such as "10/04/12" or "10-04-12" are ambiguous: 12 April 2010, 10 April 2012, or 4 October 2012. Similarly, include the name of the time zone after the time. Co-ordinated Universal Time (UTC) is preferred over Greenwich Mean Time (GMT).
6. **Have your figure panels been labelled in multiple-panel figures?** For example "(a), (b),…"**.** Have they been labelled from left to right, then top to bottom?
7. **Have all the features in the graph been described in the figure caption?**

3.3.3. Linking the figures to the text

After creating the perfect figure, students often insert the figure into their report and believe they are done. Unfortunately, the creation of the figure is not the end of the story: you need to link the content of the figure with the content of the text to help tell the story. Annotating the figure, as discussed above, is one way of helping the reader understand the figure, but the text is the other way. Consider these guidelines to describing your figure.

Make sure that if you include a figure in your text that it is a relevant figure to the story that you want to tell. The best figure does not help you get a good mark if it is not relevant and tied closely to your work.

If you include a figure, make sure that you explain it to the reader. Do not include material that is explained in the caption, but explain the content of the figure. Consider the text associated with the examples in section 3.3.1 as illustrations of how you might write your text to explain the figures.

When explaining the figures, explain the most obvious features of the figure first. Then, explain the less obvious features that may pertain to your specific conclusions. Again, annotations of the figure may help the reader easily see the same features that you are trying to describe.

Avoid the phrase "Figure 3 shows that…." Instead, place "(Figure 3)" at the end of the text that explains the figure (usually at the end of the sentence).

4. List of GEES data sources

To assist students with identifying large and useful online datasets, the following list has been assembled. The links are current as of the time of writing of this document (May 2013).

4.1. Satellite imagery and remote-sensing

Google Earth	http://www.google.co.uk/intl/en_uk/earth/index.html
Goddard Earth Sciences Data and Information Center	http://disc.sci.gsfc.nasa.gov/
NASA Giovanni: access to satellite datasets	http://disc.sci.gsfc.nasa.gov/giovanni
ESA Earth Observations datasets	https://earth.esa.int/web/guest/home
Global Land Cover Facility	http://www.landcover.org/
USGS: Global Land Cover Data	http://landcover.usgs.gov/glcc/
Landmap	http://www.landmap.ac.uk
Geostationary satellite data	http://www.ncdc.noaa.gov/gibbs
Britain from Above	http://www.britainfromabove.org.uk
Satellite data archive (need to register)	http://www.sat.dundee.ac.uk
The gateway to astronaut photography of the Earth	http://eol.jsc.nasa.gov/
Spanish aerial photos, maps, digital elevation models	http://centrodedescargas.cnig.es/CentroDescargas/buscador.do http://www2.ign.es/iberpix/visoriberpix/visorign.html
Satellite data archive (need to register)	http://www.sat.dundee.ac.uk
Sat 24	http://www.sat24.com/history.aspx

4.2. Topographic and land cover mapping

EDINA Digimap (need to check if your university is a subscriber)	http://digimap.edina.ac.uk/digimap/home
UK Land Cover Map (2007)	http://www.ceh.ac.uk/landcovermap2007.html
Global Land Cover Characterization Program (USGS)	http://landcover.usgs.gov/glcc/

4.3. Atmospheric Science and Meteorology

British Atmospheric Data Centre (need to register)	http://badc.nerc.ac.uk/data/dataset_index
Land surface climate stations (UK)	http://www.metoffice.gov.uk/research/climate/climate-monitoring/land-and-atmosphere/surface-station-records http://www.metoffice.gov.uk/climate/uk/stationdata
Sources of UK weather data	http://www.met.rdg.ac.uk/~brugge/bi.html
UK climate maps and anomalies	http://www.metoffice.gov.uk/climate/uk/anomacts
Community Data Portal, National Center for Atmospheric Research	http://cdp.ucar.edu/home/home.htm

National Climatic Data Center (US)	http://www.ncdc.noaa.gov
Archived Weather Charts	http://nomads.ncdc.noaa.gov/ncep/NCEP
Daily Weather Maps (US)	http://docs.lib.noaa.gov/rescue/dwm/data_rescue_daily_weather_maps.html
National Data Buoy Center	http://www.ndbc.noaa.gov/hmd.shtml
National Hurricane Center	http://www.nhc.noaa.gov/pastall.shtml
Plymouth State University: Make your own weather charts (mostly US)	http://vortex.plymouth.edu/u-make.html
National Center for Atmospheric Research Image Archive (US)	http://www.mmm.ucar.edu/imagearchive
Archived weather data	http://www.wunderground.com
ECMWF Archive	http://www.met.rdg.ac.uk/Data/CurrentWeather
UK and Western Europe radar imagery	http://www.meteox.co.uk/hist.aspx?URL
Surface-based cloud climatology	http://www.atmos.washington.edu/~ignatius/CloudMap
Climate of worldwide cities	http://www.weatherbase.com
University of Wyoming soundings and upper-air charts	http://weather.uwyo.edu/upperair/
Global data, web-based plotting tool	http://www.esrl.noaa.gov/psd/cgi-bin/data/getpage.pl
NCEP/NCAR Global Reanalysis Data	http://www.esrl.noaa.gov/psd/data/composites/hour/

4.4. Palaeoclimatology

NOAA – paleoclimate data (various points of access)	http://www.ncdc.noaa.gov/data-access/paleoclimate-data http://www.ncdc.noaa.gov/paleo/data.html http://www.ncdc.noaa.gov/paleo/recons.html
University of East Anglia	http://www.cru.uea.ac.uk/cru/data/paleo/

4.5. Atmospheric chemistry

UK Air Quality from DEFRA	http://uk-air.defra.gov.uk/data
Globemission	http://globemission.eu
HIAPER Pole-to-Pole Observations of Carbon Cycle and Greenhouse Gases	http://www.eol.ucar.edu/projects/hippo http://hippo.ornl.gov
World Data Centre for Greenhouse Gases	http://ds.data.jma.go.jp/gmd/wdcgg/wdcgg.html
Global SO_2 monitoring:	http://so2.gsfc.nasa.gov

4.6. Hydrology

National River Flow Archive	http://www.ceh.ac.uk/data/nrfa/
HiFlows – UK flood peak data	http://www.environment-agency.gov.uk/hiflows/91727.aspx
National Well Record Archive	http://www.bgs.ac.uk/research/groundwater/datainfo/NWRA.html
National Groundwater Level Archive	http://www.ceh.ac.uk/data/nrfa/data/groundwater_sources.html
USGS Groundwater data	http://water.usgs.gov/ogw/data.html
International Soil Moisture Network	http://ismn.geo.tuwien.ac.at/

4.7. Oceanography

National Oceanographic Data Center *Specific datasets can be searched for using the following link:*	http://www.nodc.noaa.gov/index.html http://data.nodc.noaa.gov/geoportal/catalog/search/search.page
Bathymetry	http://www.geomapapp.org
TAO Buoy Array in Tropical Pacific Ocean	http://www.pmel.noaa.gov/tao/disdel/disdel.html

4.8. Glaciology

National Snow and Ice Data Center	http://nsidc.org/data/
World Data Center for Glaciology	http://www.wdcgc.spri.cam.ac.uk/
Circumpolar Active Layer Monitoring	http://www.gwu.edu/~calm/

4.9. Environmental Science

Environment Agency: My Environment	http://www.myenvironment.org.uk/
Environment Agency: Data & statistics	http://www.environment-agency.gov.uk/research/library/data/default.aspx
Environment Agency: What's in your backyard?	http://maps.environment-agency.gov.uk/wiyby
UK Radioactive Incident Monitoring Network	http://archive.defra.gov.uk/evidence/statistics/environment/radioact/radrimnet.htm
Natural Resources Conservation Service Soils website	http://soils.usda.gov/
USGS Water Quality information	http://water.usgs.gov/owq/

European Chemicals Agency	http://echa.europa.eu/web/guest/information-on-chemicals/registered-substances

4.10. Geology & Geophysics

BGS OpenGeoscience *Provides access to a broad range of datasets including GIS datasets, borehole data and aerial photographs.*	http://www.bgs.ac.uk/opengeoscience/
USGS *Provides access to a broad range of maps, images and other datasets.*	http://www.usgs.gov/pubprod/
UK Onshore Geophysical Library	http://maps.lynxinfo.co.uk/UKOGL_LIVEV2/main.html
Virtual Seismic Atlas	http://www.seismicatlas.org/
KML Geology (San Diego State University) *Provides access to a range of geoscience visuals designed to be viewed in Google Earth.*	http://www.geology.sdsu.edu/kmlgeology/
Geochemistry of Rocks of the Oceans and Continents	http://georoc.mpch-mainz.gwdg.de/georoc/Start.asp
North American Volcanic and Intrusive Rock Database	http://www.navdat.org/index.cfm
Petrological Database	http://www.earthchem.org/petdb
EarthChem portal (access to multiple datasets)	http://www.earthchem.org/portal
Paleobiology Database	http://www.pbdb.org

4.11. Extraterrestrial geological sciences

Meteorite classification database	http://www.lpi.usra.edu/meteor/
Meteorite chemistry database	http://www.metbase.de/home.html
Apollo sample database	http://www.lpi.usra.edu/lunar/samples/atlas/
Apollo sample compendium database	http://curator.jsc.nasa.gov/lunar/compendium.cfm
Lunar image database	http://www.lpi.usra.edu/lunar/lunar_images/
Planetary Data System – (all archived released NASA image datasets)	http://pds.nasa.gov
Map a Planet– planetary image end user tool	http://www.mapaplanet.org
Lunar Impact Crater Catalogue	http://www.lpi.usra.edu/lunar/surface/Lunar_Impact_Crater_Database_v24May2011.xls

4.12. Other tools

MatLab mapping package	http://www.eos.ubc.ca/~rich/map.html

5. Acknowledgements

We thank the following individuals from the Keele University and the University of Manchester who provided suggestions and web sites for this report: James Allan, Grant Allen, Victoria Egerton, Jonathan Fairman, Sam Illingworth, Katherine Joy, Francis Livens, Ian Lyon, Julian Mecklenburgh, Neil Mitchell, Jennifer Muller, Alison Pawley, Ian Oliver, Thomas Seers and Ian Stimpson. We thank Helen Walkington, Discipline Lead for Geography, Earth and Environmental Sciences of the Higher Education Academy for her support and comments to improve this report.

6. References

Anscombe, F. J. (1973) Graphs in statistical analysis. *The American Statistician* 27 (1), 17–21.

Banacos, P. C. (2011) Box and whisker plots for local climate datasets: Interpretation and creation using Excel 2007/2010 [online]. National Oceanic and Atmospheric Administration, National Weather Service, Eastern Region Technical Attachment 2011-01, 20 pp. Available from: http://docs.lib.noaa.gov/noaa_documents/NWS/NWS_ER/Eastern_Region_Tech_Attachment/TA_2011-01.pdf. [accessed 02/03/2013]

Brennan, J., Johnstone, B., Little, B., Shah,T. and Woodley,A. (2001). The Employment of UK Graduates: comparisons with Europe and Japan [online]. Report to the HEFCE by the Centre for Higher Education Research and Information and The Open University Bristol. Available from: http://www.hefce.ac.uk/Pubs/hefce/2001/01_38.htm [accessed 02/03/2013]

Foreman, B. Z., Heller, P. L., Clementz, M. T. (2012) Fluvial response to abrupt global warming at the Palaeocene/Eocene boundary. *Nature,* 491**,** 92–95.

McCullough, B. D. (2008) Editorial: Special section on Microsoft Excel 2007. *Computational Statistics and Data Analysis,* 52, 4568–4569.

Schultz, D. M., (2009) Eloquent Science: A Practical Guide to Becoming a Better Writer, Speaker, and Atmospheric Scientist [online]. *Bulletin of the American Meteorological Society,* 440 pp. Excerpts and resources available from: http://www.eloquentscience.com [accessed 02/03/2013]

SCONUL (2011) *The SCONUL Seven Pillars of Information Literacy* [online]. SCONUL Working Group on Information Literacy. Available from: http://www.sconul.ac.uk/sites/default/files/documents/coremodel.pdf [accessed 02/03/2013]

Symonds, R. B., Rose, W.I., Bluth, G., Gerlach, T.M. (1994) Volcanic gas studies: methods, results, and applications, in Carroll, M.R., and Holloway, J.R. (eds.). *Volatiles in Magmas: Mineralogical Society of America Reviews in Mineralogy.* 30, 1–66.

Thomas, C. (2008) An employer's perspective on the recruitment and retention of GEES graduates in the Environmental Sector. *Planet,* 19**,** 44–46.

Verbout, S. M., Brooks, H. E., Leslie, L. M., Schultz, D. M. (2006) Evolution of the U.S. tornado database: 1954-2004. *Weather Forecasting,* 21, 86-93.

Wu, B., Handorf, D., Dethloff, K., Rinke, A., Hu. A. (2013) Winter weather patterns over northern Eurasia and Arctic sea ice loss. *Monthly Weather Review.* Available from: http://journals.ametsoc.org/doi/abs/10.1175/MWR-D-13-00046.1?journalCode=mwre[accessed 02/03/2013]

Contact us

The Higher Education Academy
Innovation Way
York Science Park
Heslington
York
YO10 5BR

+44 (0)1904 717500
enquiries@heacademy.ac.uk

ISBN: 978 -1-907207-93-8

The Higher Education Academy (HEA) is a national body for learning and teaching in higher education. We work with universities and other higher education providers to bring about change in learning and teaching. We do this to improve the experience that students have while they are studying, and to support and develop those who teach them. Our activities focus on rewarding and recognising excellence in teaching, bringing together people and resources to research and share best practice, and by helping to influence, shape and implement policy - locally, nationally, and internationally. The HEA supports staff in higher education throughout their careers, from those who are new to teaching through to senior management. We offer services at a generic learning and teaching level as well as in 28 different disciplines. Through our partnership managers we work directly with HE providers to understand individual circumstances and priorities, and bring together resources to meet them. The HEA has knowledge, experience and expertise in higher education. Our service and product range is broader than any other competitor.

www.heacademy.ac.uk | www.twitter.com/heacademy

To request copies of this report in large print or in a different format, please contact the communications office at the Higher Education Academy: 01904 717500 or pressoffice@heacademy.ac.uk