The UKCP09 Outputs and Metadata Specification

Release 1.1

02 January 2014

British Atmospheric Data Centre,
UK Climate Impacts Programme

Version History

Version Number	Who?	Date	Notes
0.1	Ag Stephens (AS)	10/06/2009	Compiled previous documents by AS and Paul Bowyer (PB) of UKCIP
1.0	AS	15/06/2009	Version 1.0 released.
1.1	AS	02/01/2014	Updated links to Met Office domain.

Contents

1. Introduction

1.1. Purpose of this document

The UKCP09 Project is a collaboration involving scientists, communicators, government departments, data experts and technologists. The User Interface (UI) provides the access to project outputs. These outputs must also be consistent and well-described with an appropriate level of information (metadata) passed to users alongside the data.

This document explains the output formats delivered by the UI. Particular attention is paid to the formatting of CSV outputs which are likely to be the most commonly accessed by our users.

1.2. Context

1.3. Approach taken in UKCP09

1.3.1. Requirements

Requirements were gathered from the various stakeholders with particular engagement with the UKCP09 Users' Panel. It was established that the following were important:

- a standard text format for UKCP09 outputs
- a standard way of handling gridded UKCP09 data using this text format
- a standard binary format for UKCP09 outputs
- a standard GIS format for UKCP09 outputs[1]
- a standard way of handling metadata
- a standard way of handling large/multiple file outputs

[1] Note that at present the option to download shapefiles has been disabled on the UI. The project management group are further considering the suitability of this format for mapped outputs that are probabilistic in nature.

1.3.2. Format design principles

Based on our understanding of the requirement we developed an approach built on the following principles:

- We should build on existing standard formats rather than making up new formats that are not tried and tested
- We must provide internal conversions between the different formats
- We should provide clear guidance to UKCP09 users via a detailed format specification document, including examples.
- We should ensure that metadata always travels with the outputs

1.3.3. Summary of our approach

We decided that a comma-separated variable (CSV) format was a most appropriate text format as it is human-readable, machine-readable and readily usable on any computing platform. As a result of this decision we have developed tools to convert a range of different data structures into 1D or 2D arrays so that they can easily be displayed in a table.

As a leader in marine and climate science CF-netCDF was most appropriate as a binary format. In all cases (including image outputs) it was decided that a metadata file should be delivered along with the data products.

In terms of handling multiple and large files, we agreed that grouping outputs into zip files was most appropriate as it would reduce download wait times and the need to make multiple download actions. Finally, it was decided that Weather Generator outputs should not be post-processed as users may be familiar with their formats from other projects.

2. Note on UKCP09 data licence

All users of the UKCP09 UI must agree to the Terms and Conditions of use of located at:

http://ukclimateprojections.metoffice.gov.uk/23476

3. UKCP09 output formats

3.1.1. Test format: CSV

3.1.1.1. What is a CSV format and why is it useful?

Text files are considered the most desirable format by the UKCP09 user community. They have the following advantages:

- they are human-readable
- they are machine-readable
- they can be read without specialist software

The main drawback with a text-based format is that files are larger in size than an equivalent binary format. We overcome this disadvantage by delivering all outputs in zip files, which can typically reduce overall volume by a factor of 5 – 10 times.

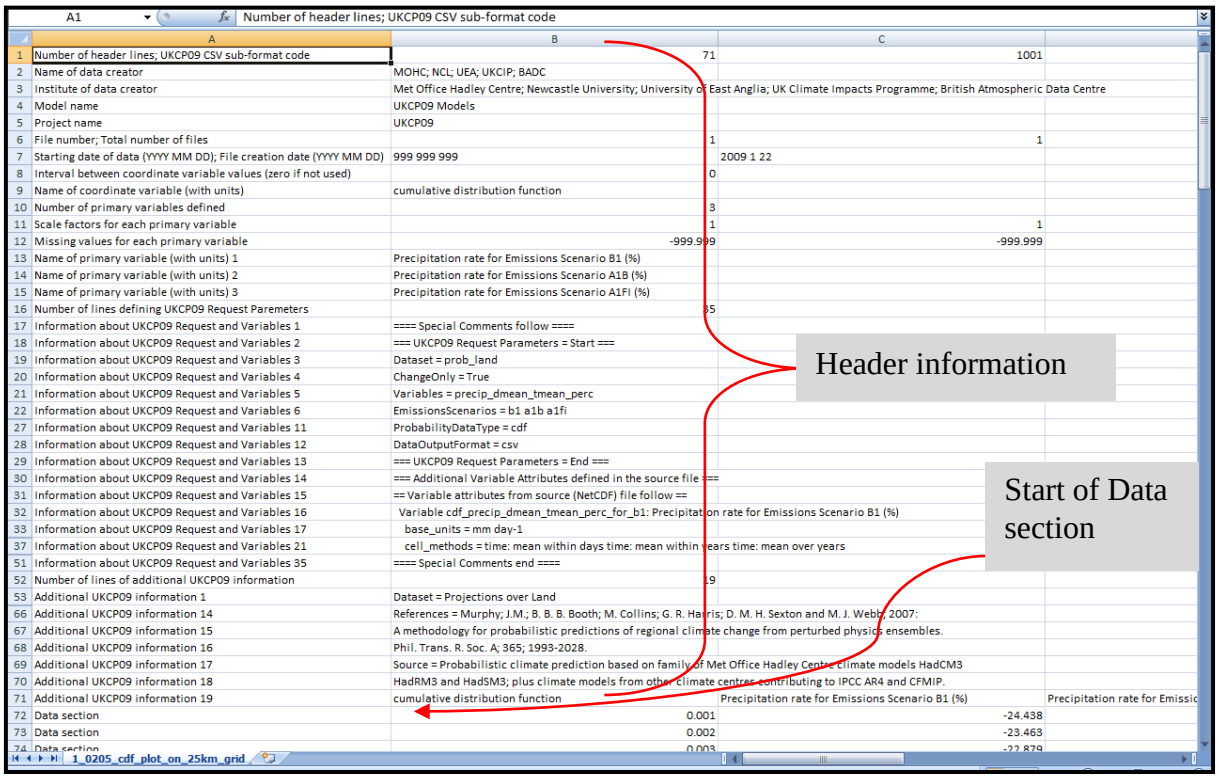
CSV files are a particular form of text file where each row contains a set of fields separated by commas. The extra advantages of CSV files are:

- additional structure is imposed by tabulated the data into rows and columns
- spreadsheet packages (such as Microsoft Excel) can read CSV without any modification
- writing software to process CSV files is straightforward

However, CSV as a format still lacks any detailed structure. We have chosen to build upon this to generate a format that includes specific sections for metadata and annotations as well as the main data section.

3.1.1.2. The UKCP09 CSV format

All of the UKCP09 CSV output files are structured in the same way. They consist of a header section containing a number of lines, with details of how the data were generated (emissions scenario etc). This header information (metadata) is then followed by a data section which contains the actual data. This structure is illustrated in figure 1. Whilst there are differences in the way the data section is structured between the different output types, the header structure is consistent between output types.



	A	B	C	
1	Number of header lines; UKCP09 CSV sub-format code	71	1001	
2	Name of data creator	MOHC; NCL; UEA; UKCIP; BADC		
3	Institute of data creator	Met Office Hadley Centre; Newcastle University; University of East Anglia; UK Climate Impacts Programme; British Atmospheric Data Centre		
4	Model name	UKCP09 Models		
5	Project name	UKCP09		
6	File number; Total number of files	1	1	
7	Starting date of data (YYYY MM DD); File creation date (YYYY MM DD)	999 999 999	2009 1 22	
8	Interval between coordinate variable values (zero if not used)	0		
9	Name of coordinate variable (with units)	cumulative distribution function		
10	Number of primary variables defined	3		
11	Scale factors for each primary variable	1	1	
12	Missing values for each primary variable	-999.999	-999.999	
13	Name of primary variable (with units) 1	Precipitation rate for Emissions Scenario B1 (%)		
14	Name of primary variable (with units) 2	Precipitation rate for Emissions Scenario A1B (%)		
15	Name of primary variable (with units) 3	Precipitation rate for Emissions Scenario A1FI (%)		
16	Number of lines defining UKCP09 Request Paremeters	35		
17	Information about UKCP09 Request and Variables 1	==== Special Comments follow ====		
18	Information about UKCP09 Request and Variables 2	=== UKCP09 Request Parameters = Start ===		
19	Information about UKCP09 Request and Variables 3	Dataset = prob_land		
20	Information about UKCP09 Request and Variables 4	ChangeOnly = True		
21	Information about UKCP09 Request and Variables 5	Variables = precip_dmean_tmean_perc		
22	Information about UKCP09 Request and Variables 6	EmissionsScenarios = b1 a1b a1fi		
27	Information about UKCP09 Request and Variables 11	ProbabilityDataType = cdf		
28	Information about UKCP09 Request and Variables 12	DataOutputFormat = csv		
29	Information about UKCP09 Request and Variables 13	=== UKCP09 Request Parameters = End ===		
30	Information about UKCP09 Request and Variables 14	=== Additional Variable Attributes defined in the source file ===		
31	Information about UKCP09 Request and Variables 15	== Variable attributes from source (NetCDF) file follow ==		
32	Information about UKCP09 Request and Variables 16	Variable cdf_precip_dmean_tmean_perc_for_b1: Precipitation rate for Emissions Scenario B1 (%)		
33	Information about UKCP09 Request and Variables 17	base_units = mm day-1		
37	Information about UKCP09 Request and Variables 21	cell_methods = time: mean within days time: mean within years time: mean over years		
51	Information about UKCP09 Request and Variables 35	==== Special Comments end ====		
52	Number of lines of additional UKCP09 information	19		
53	Additional UKCP09 information 1	Dataset = Projections over Land		
66	Additional UKCP09 information 14	References = Murphy; J.M.; B. B. B. Booth; M. Collins; G. R. Harris; D. M. H. Sexton and M. J. Webb; 2007:		
67	Additional UKCP09 information 15	A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles.		
68	Additional UKCP09 information 16	Phil. Trans. R. Soc. A; 365; 1993-2028.		
69	Additional UKCP09 information 17	Source = Probabilistic climate prediction based on family of Met Office Hadley Centre climate models HadCM3		
70	Additional UKCP09 information 18	HadRM3 and HadSM3; plus climate models from other climate centres contributing to IPCC AR4 and CFMIP.		
71	Additional UKCP09 information 19	cumulative distribution function	Precipitation rate for Emissions Scenario B1 (%)	Precipitation rate for Emissic
72	Data section	0.001	-24.438	
73	Data section	0.002	-23.463	
74	Data section	0.003	-22.879	

Figure 1. General structure of the comma-separated-values (CSV) files in the data, as viewed in Microsoft Excel 2007.

The first column explains each line
The first column of the file (i.e. before the first comma on each line) provides an explanation of the content of the rest of the line. This column is included to make it easier to read the format. Many users will choose to hide or remove this column once accustomed to working with the outputs. This can be done in a spreadsheet package or programmatically.

CSV file header
The CSV header consists of three main sections:

i. General characteristics of the file the data contained therein
ii. Information about the UKCP09 request parameters used to generate a particular output
iii. Additional information about UKCP09 and the data sources

i. General Characteristics
The general characteristics of the data and the file are provided in the header of all UKCP09 CSV outputs. Most of those listed below are always present but some are dependent on the sub-format (as defined in the first line):

- **Number of header lines and UKCP09 CSV sub-format code**: a number defining the number of lines of header information, this is the number of lines before the data section is reached; and a code defining the exact format that the data are stored in. These are based on the NASA Ames sub-formats (documented at: http://badc.nerc.ac.uk/help/formats/NASA-Ames).
- **Name of data creator**: text detailing the name of the data creator, this will read MOHC (for Met Office Hadley Centre), NCL (for Newcastle University), UEA (University of East Anglia), UKCIP (for UK Climate Impacts Programme), BADC (for British Atmospheric Data Centre).
- **Institute of data creator**: see data creator information above.
- **Model name**: this will read UKCP09 models.
- **Project name**: name of the project, which will read UKCP09.
- **File number and total number of files**: the file number in the download, and the total number of files that have been produced to provide the output requested.
- **Starting date of the data, and file creation date**: dates will be in YYYY MM DD format, with two entries – one for the start date of the data, and the other for the file creation date. Note that in some cases the first value is populated with "999 999 999". This indicates that the time information is available elsewhere in the file.
- **Interval between coordinate variable values**: one or more numeric values that will tally with the number of coordinate variables there are in the data file, which defines the numeric interval between the coordinate variable values.
- **Length of coordinate variables**: a number indicating the length of the coordinate variables in the header, e.g. if the coordinate variables were latitude and longitude, these numbers would correspond to the number of values for each coordinate variable.
- **Number of coordinate variable values explicitly defined in header**: the number of coordinate variable values defined in the header, e.g. if the coordinate variables were latitude and longitude, the number would be 2.

- **Values of coordinate variable**: there can be multiple entries of this item, which provides a list of the values of the coordinate variables which corresponds to the total number given as being explicitly defined in the header.
- **Name of coordinate variable (with units)**: there can be multiple entries of this item, which gives the name of the coordinate variable and associated units.
- **Number of primary variables defined**: the number of primary variables defined in the file.
- **Scale factors for each primary variable**: a number defining the scale factor to be applied to the data values. A value of one means that there is no scaling, and that the numbers given in the data section of the file can be used as is, without needing to be further manipulated. The number of instances of this value will correspond to the number of primary variables.
- **Missing values for each primary variable**: if there are missing values, this will be indicated by the appearance of the number -999.999. The number of instances of this value in the header will correspond to the number of primary variables.
- **Name of primary variable (with units)**: there can be multiple entries for this item, and each describes the name of the primary variable. The primary variable corresponds to the climate variable.
- **Number of auxiliary variables:** auxiliary variables are not used in UKCP09 outputs.

After this section, there is a single line entry which defines how many lines of i*nformation about UKCP09 request and variables* follow.

ii. Information about UKCP09 request and variables

Following the first lines of text described above, there follows a section which provides further information on how the primary variables were extracted, and which scenario settings they were generated from. These are listed as Information about UKCP09 Request and Variables in the CSV files. The number of instances of this entry tallies with the number of lines of UKCP09 request parameters, so that if there were 12 lines of UKCP09 request parameters, the entries in the CSV file would be listed as:

Information about UKCP09 Request and Variables 1
Information about UKCP09 Request and Variables 2
.
.
Information about UKCP09 Request and Variables 12

This part of the header is split into two sub-sections (which are easily visible in the CSV files), the first of which contains information on the UKCP09 request parameters, whilst the second sub-section contains information on additional variable attributes. Summaries of these two sub-sections are given below, in tables 1 and 2, respectively.

Table 1. Summary description of the full suite of possible UKCP09 request parameters, in the header of the CSV files.

Request parameter	Description	Possible attribute values in the data
Dataset	Defines what the data source is.	prob_land prob_marine storm_surge sea_level_rise

ChangeOnly	Climate change type. UKCP09 provides information relating to both absolute climate change, and relative to a 1961-1990 baseline	True
Variables	The climate variable to which the numbers in the CSV file relates.	A long list in some cases, and a full description can be found in the Additional Variable Attributes section of the header.
EmissionsScenarios	Describes which of the IPCC SRES emissions scenarios the data were generated from or relate to. The values correspond to the low, medium, and high emissions scenarios, in UKCP09.	B1 (low) A1B (medium) A1FI (high)
TimePeriods	The time period(s) to which the data apply. Some of the time slices may not be available for certain products (output types). The time periods (in UKCP09 climate change projections are provided for seven time periods), are numbered according to the number of days since 1 December 2009, on the basis of 360 days in the year. Thus 5400 represents the 2020s (5400/360 is 15 years, which is 2024 and thus the 2020s time period), 9000 (the 2030s), 12600 (2040s), 16200 (2050s), 19800 (2060s), 23400 (2070s), 27000 (2080s).	2010-2039 2020-2049 2030-2059 2040-2069 2050-2079 2060-2089 2070-2099
TemporalAverages	The period over which the data have been averaged.	jja: June, July, August (summer) djf: December, January, February (winter) son: September, October, November (autumn) mam: March, April, May (spring) jan: January jul: July
LocationType	Describes the spatial unit to which the data relate.	Grid_box_25km Marine Region
Location	Provides an identifier for the location type e.g. region = Wales.	Either a Grid Box ID or the name of an aggregated area.
ProbabilityDataType	Probabilistic model output can be presented as a	pdf cdf

	probability density function (pdf), or a cumulative probability distribution (cdf), or the sampled model data (samp_data).	samp_data
DataOutputFormat	Output format of the file.	csv – a comma-separated-values file
BBox	The bounding box defining the spatial extent that the data relates to, a four value vector [x1,y1,x2,y2]	x1: longitude of lower left y1: latitude of lower left x2: longitude of upper right y2: latitude of upper right
Percentiles	The probabilistic data is produced with or at a given probability level, at or below which a certain proportion of the total number of probabilistic climate change projections occur.	0.1 – 99.9
SamplingMethod	There are various ways in which the modelled data can be sampled. In the data the modelled data has been sampled at random.	select_all, random, by_id, by_percentile
NumberOfRandomSamples	When random sampling the number of samples is defined here.	100 - 10000
SampleIdList	A list of IDs representing samples (or model variants)	List of numbers of length 100 - 10000
SamplingVariable1	When sampling by percentile, this is the first variable to sample by.	A climate variable name
SamplingTemporal AveragingPeriod1	When sampling by percentile, this is the first temporal average to sample by.	A temporal average (see above)
SamplingPercentile1	When sampling by percentile, this is the first percentile to sample by.	A percentile values in the range inclusive of 5 - 95
SamplingVariable2	When sampling by percentile, this is the second variable to sample by.	A climate variable name
SamplingTemporal AveragingPeriod2	When sampling by percentile, this is the second temporal average to sample by.	A temporal average (see above)
SamplingPercentile2	When sampling by percentile, this is the second percentile to sample by.	A percentile values in the range inclusive of 5 - 95

Table 2. Summary description of the full suite of possible Additional Variable Attributes, in the UKCP09 data set.

Additional Variable Attribute	**Description**

variable	Describes in full what the climate variable is.
units	Defines the unit of measurement for the climate variable.
base_units	Defines the unit of measurement for the absolute value of the climate variable. For example, when precipitation is selected for future climate change then it is represented as a percentage change from the baseline climate. The unit for the variable is % but the base unit is mm/day.
long_name	Provides a long name description of what the climate variable is.
grid_mapping	This defines on what kind of grid the data relate to, which is useful when considering the latitude and longitude data associated with the map plot output types (products). This is typically the rotated grid associated with the MOHC climate model. Regular lat/lon gridded data does not have this attribute
coordinates	This is a text descriptor of the spatial information used to extract the data from the full database.
comment	Some additional text to provide further clarity as to what the climate variable actually is.
cell_methods	Some text to describe the temporal average or other processing that has gone on with relation to a coordinate variable.
source	A reference to a source document that might help a user find out some more about the particular climate variable or data set.

After this section, there is a single line entry which defines how many lines of *additional UKCP09 information* follow.

iii. Additional UKCP09 information

This part of the header is common to all the UKCP09 data files, and contains mostly identical information in each file, apart from certain entries that relate to each data source. The entries from a typical output are provided in table 3.

Table 3. Structure and detail of the Additional UKCP09 Information header. As these are data some of the information text will be placeholders, and have fictional names.

Additional UKCP09 information	**Topic**	**Information text**
Number of lines of additional UKCP09 information	This section of the file*	21
Additional UKCP09 information 1	Data format	Format_Details = This CSV format header is based on the NASA Ames data exchange format.
Additional UKCP09 information 2		The UKCP09 format specification document is available at:
Additional UKCP09 information 3		http://ukclimateprojections-ui.metoffice.gov.uk/ui/docs/formats/ukcp09_file_format_spec.pdf
Additional UKCP09	Data source*	Dataset = UK probabilistic projections of climate change over land

information 4		
Additional UKCP09 information 5	File creation log*	History = File generated on 2009-06-16 17:06:34 by UKCP09 User Interface (v1.0)
Additional UKCP09 information 6	UKCP09 Data Licence	Licence_Details = All products generated by the UKCP09 User Interface are licensed under the Terms and Conditions at:
Additional UKCP09 information 7		http://ukclimateprojections.metoffice.gov.uk/23476
Additional UKCP09 information 8		The licence granted to the Licensee operates as a permission only and does not imply any obligation
Additional UKCP09 information 9		or liability or any conditions warranties or representations in relation to the accuracy of the Data
		on the part of the Department and organisations responsible for producing UKCP09.
Additional UKCP09 information 11		The Licensee acknowledges that the Department has made the Data available on the basis that they reflect
Additional UKCP09 information 12		current scientific understanding and that the Licensee is aware of the uncertainty described in the accompanying reports
Additional UKCP09 information 13		Available on the UK Climate Projections web site.
Additional UKCP09 information 14	Locations and grids	Location_Details = The grids and spatial averages used in UKCP09 are documented at:
Additional UKCP09 information 15		http://ukclimateprojections-ui.metoffice.gov.uk/ui/docs/grids
Additional UKCP09 information 16		The actual locations of grid box IDs are available on the grids pages.
Additional UKCP09 information 17	References	References = Murphy; J.M.; B. B. B. Booth; M. Collins; G. R. Harris; D. M. H. Sexton and M. J. Webb; 2007:
Additional UKCP09 information 18		A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles.
Additional UKCP09 information		Phil. Trans. R. Soc. A; 365; 1993-2028.

19		
Additional UKCP09 information 20	Data source*	Source = Probabilistic climate prediction based on family of Met Office Hadley Centre climate models HadCM3
Additional UKCP09 information 21		HadRM3 and HadSM3; plus climate models from other climate centres contributing to IPCC AR4 and CFMIP.
Additional UKCP09 information 22	Header line for Data section	cumulative distribution function, Precipitation rate for Emissions Scenario A1FI (%)

* These values will differ depending on output type and data source.

CSV data section

Most of the UKCP09 datasets are stored in data structures of 3-6 dimensions. CSV are most useful when represented data in table form. This means that all data outputs have to be *squeezed* from their original structure into a 1- or 2- dimensional structure. Details of the how we have re-structured the data for each output type and data source is provided in appendix 2.

The first field in each row contains the text “Data section”.

When the UKCP09 sub-format (given in the first line of the header) is 1001 then the last line of the *additional UKCP09 information* (as shown in table 3) includes a header for each column of the Data Section.

The structure of the data section depends on the sub-format. The different sub-formats are structured as follows:

Sub-format 1001

Sub-format 1001 is used to provide one or more variables that are defined against one dimension. The values of the dimension are provided in the second column. The third column contains the first variable. If more than one variable is defined in the file then additional columns are added for each variable. The data section is laid out as in the example in table 4.

Table 4. Example sub-format 1001 snippet for two variables defined against one dimension

Additional UKCP09 information 22	**ID of sample representing model variant**	**Mean air temperature at 1.5m for DJF (degC)**	**Precipitation rate for DJF (%)**
Data section	0	1.167	32.388
Data section	1	2.137	18.016
Data section	2	2.241	9.63
Data section	3	1.994	27.19
Data section	4	3.405	29.012

Sub-format 3010

Sub-format 3010 is used specifically to represent 2-dimensional gridded data outputs. It allows the x- and y-axes of the data structure to be represented in a table that is easy to read visually (see figure 2) or to read into plotting packages.

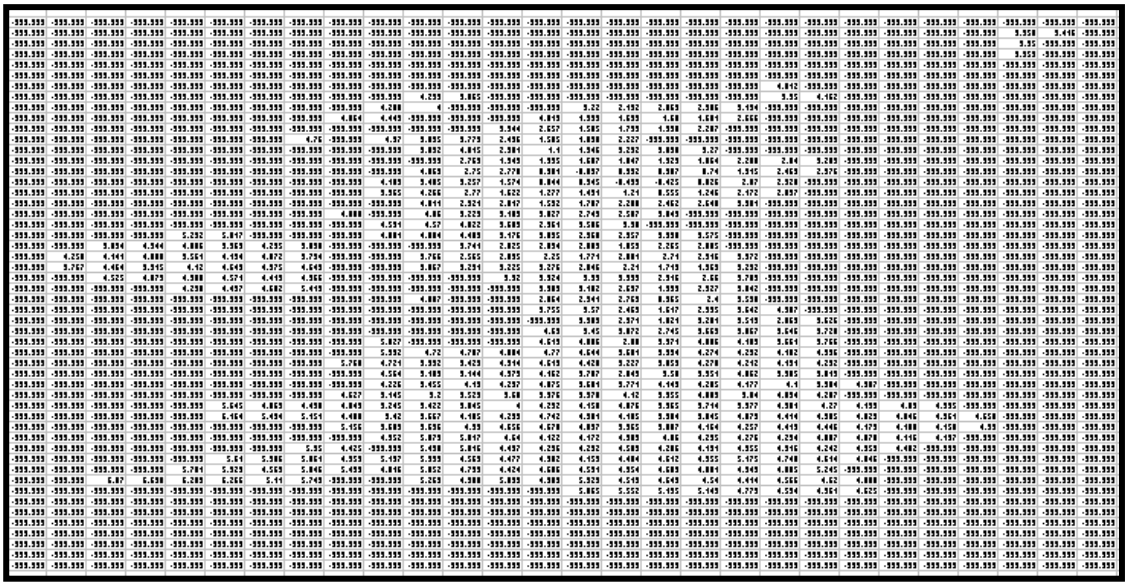


Figure 2. When viewing the 2D array of gridded data it is possible to zoom out and identify known geographical areas because of the land/sea mask.

Missing values are defined in the header but are typically defined as -999.999. The data section is laid out as in the example in table 5. Note that the column headers are not included for sub-format 3010. The first line of the data section provides a single value for a dimension that the entire grid is defined against. The example in table 5 represents a time value of 5400. This is relative to the units of time given in the header of "days since 2009-12-01 00:00:00". This equates to the central period of the Time Period 2010-2039 and so is representative of the 2020s.

Table 5. Example sub-format 3010 snippet for two variables defined against one dimension

Data section	5400				
Data section	4.449	-999.999	-999.999	-999.999	4.019
Data section	-999.999	-999.999	-999.999	3.944	2.657
Data section	4.37	3.895	3.779	2.436	1.585
Data section	-999.999	3.832	4.015	2.301	1.1
Data section	-999.999	-999.999	2.769	1.343	1.395
Data section	-999.999	4.069	2.75	2.778	0.981
Data section	4.183	3.405	3.257	1.574	0.844
Data section	3.365	4.266	2.77	1.622	1.277

Further details on how to interpret spatial information is available in chapter 7.

For details of the grids that UKCP09 outputs are defined against please see the grids documentation pages at:

http://ukclimateprojections-ui.metoffice.gov.uk/ui/docs/grids/

3.1.1.3. Software for reading CSV files

Text editors

Any simple text editor on any will be able to open and view a CSV file. Example utilities are *Notepad* on Windows and *nedit* on Unix/Linux.

Spreadsheet packages

Spreadsheet packages are able read (and write) CSV files. They also format the columns making it significantly easier to read the file than in a text editor. Examples of spreadsheet packages are Microsoft Excel (Windows) and OpenOffice (Unix/Linux and Windows).

Reading CSV with your own programmes

Note that the UKCP09 CSV format has been specifically designed so that it can be read programmatically. Following the description in this document it should be relatively simple to write code in any language to read and manipulate the UI outputs.

Note that the first column of the file represents only metadata about the content of each line. In writing software to read the files it would be straightforward to omit these lines if required.

3.2. Binary format: CF-netCDF

3.2.1. Why provide a binary format

Text formats are considered a core requirement but for managing the data in the archive, efficiency and appropriate handling of large data volumes we also require a binary format. The CF-netCDF is the chosen format given its common usage in climate sciences, strong software support and detailed metadata conventions.

3.2.2. What is NetCDF format?

NetCDF (Network Common Data Form) is both a data format and a set of libraries to read and write the format in many commonly used programming languages. It has become a *de-facto* standard in the climate and marine science communities due to its simplicity of use, robust software and portability. As a result there are numerous software packages that allow manipulation and visualisation of data written in netCDF.

Note that **netCDF version 3.3** is used in UKCP09. More recent versions are not yet commonly supported by software tools so 3.3 was chosen.

3.2.3. What is the structure of a NetCDF file?

A NetCDF file is made up of three basic components:

- variables
- dimensions
- attributes

The variables store the actual data, the dimensions give the relevant dimension (axis) information for the variables, and the attributes provide auxiliary information about the variables or the dataset itself.

This structure is flexible enough to describe all the data used in the UKCP09 project. Please refer to appendix 2 for details of how some variables are *squeezed* before being written to netCDF output files.

More information about the netCDF format is available from the Unidata web site at:

http://www.unidata.ucar.edu/netcdf

3.2.4. The CF Metadata Conventions for NetCDF

The Climate and Forecasts (CF) Metadata Conventions are guidelines and recommendations as to where to put information within a netCDF file. They also provide advice as to what type of information you might want to include. CF conventions allow the creator of the dataset to include information about the data and the dataset itself (metadata) in a structured way, which makes it easier for other users to retrieve the information.

The project makes use of the CF Conventions to improve the quality and structure of the metadata held in the data files.

More information about the CF Conventions is available at:

http://cf-pcmdi.llnl.gov/

3.2.5. Pros and cons of CF-netCDF

From the perspective of the UKCP09 project the main selling points for CF-netCDF are:

- it is efficient to store, read and write
- it can be used throughout the UI data archive and data management system
- it is a standard for climate/ocean/atmosphere users of binary data
- large volumes are far easier to handle in binary
- it is a metadata-rich format (when adopting the CF Conventions)
- it can handle multi-dimensional arrays from 1-D to 6-D[2]

[2] Some of the UKCP09 datasets are stored in a 6-dimensional data structure.

- the software libraries are all freely available
- many higher-level software packages handle NetCDF (and many of them are free!)

The drawbacks of NetCDF –CF are:

- it is not directly human-readable
- it is less compact than some binary formats
- in spite of standardisation efforts it is still possible to produce "messy" NetCDF – but adherence to CF helps overcome this

3.2.6. Software for reading NetCDF files

NetCDF is supported in all major languages and operating systems (including Windows). The Unidata web site documents many of the available libraries and software utilities freely available at:

http://www.unidata.ucar.edu/software/netcdf/

3.3. Grouping and compressing: Zip files

3.3.1. What is a zip file?

The **ZIP** file format is a data compression and archive format. A zip file contains one or more files that have been compressed to reduce file size. Zip files are can be read on all operating systems. They are appropriate for UKCP09 outputs because:

i. they provide a significant reduction in file volumes: reducing download times and storage costs
ii. they provide a mechanism for packaging up a group of files into a single file: allowing multiple outputs to be downloaded in a single transaction

3.3.2. Why use zip files?

Versions of Microsoft Excel in general usage typically have a limit of 256 columns by 64000 rows. Excel 2007 has limits of 16,000 columns and 1 million rows but we anticipate most users to be using older versions for the foreseeable future.

As a consequence of these limitations we are dividing many outputs into multiple files. The main example is the Weather Generator outputs that produce 100s - 1000s of output files. These can be grouped into a small number of zip files for download.

Text files are very inefficient in terms of storage. Using zip files typically reduces file sizes by 5 – 10 times. This also means that they can be downloaded quicker.

3.3.3. Handling zip files

When you have downloaded a zip file you can usually extract the contents of the file by double-clicking on the file.

3.3.3.1. On Windows operating systems

In the unlikely event that your system does not have an unzip utility you can download free utilities. These can be easily listed by typing “free zip utility windows” into a search engine.

3.3.3.2. On Unix/Linux operating systems

On the Unix/Linux operating systems you may prefer to work at the command line. Zip files can be unzipped using the *unzip* command, simply by typing:

```
unzip <your_file_name>
```

4. Handling metadata in UKCP09 outputs

4.1. Importance of preserving metadata

The importance of preserving metadata along with data outputs cannot be overstated. A user of UKCP09 outputs might want to ask scientific questions about the outputs, they may wish to include the outputs in scientific publications, or raise a query about the data content. In any of these instances it is vital that they have detailed metadata to identify the UI request that was made in full. The metadata should therefore be rich enough to allow reproduction of a given output and also provide sufficient information about the derivation of the data. This requirement applies to both raw data and image outputs.

4.2. The request metadata file

All outputs are accompanied by an XML file that contains details of their request. This is provided in the zip file downloaded by the user regardless of whether a plot or data output was requested. Figure 3 shows an example of the metadata file.

```
<UKCP09RequestDetails>
    <Dataset>prob_land</Dataset>
    <ChangeOnly>True</ChangeOnly>
    <Variables>spechum_dmean_tmean_perc</Variables>
    <EmissionsScenarios>a1b</EmissionsScenarios>
    <TimePeriods>2020-2049</TimePeriods>
    <TemporalAverages>son</TemporalAverages>
    <SpatialAverage>grid_box_25km</SpatialAverage>
    <BBox>-15.94 48.99 7.44 61.63</BBox>
    <Percentiles>50.0</Percentiles>
    <ProbabilityDataType>cdf</ProbabilityDataType>
    <FontSize>large</FontSize>
    <ImageOutputFormat>png</ImageOutputFormat>
    <DrawValues>False</DrawValues>
    <ShowRegions>False</ShowRegions>
    <ShowRiverBasins>True</ShowRiverBasins>
    <Thumbnail>False</Thumbnail>
    <ReUseURL>http://ukclimateprojections-ui.metoffice.gov.uk/ui/
submit/submit.php?submitid=6873124523912373654 75xF</ReUseURL>
</UKCP09RequestDetails>
```

Figure 3. Example metadata XML file.

4.3. Metadata on image outputs

All image outputs also have a metadata section above the plot. Figure 4 shows a simple example.

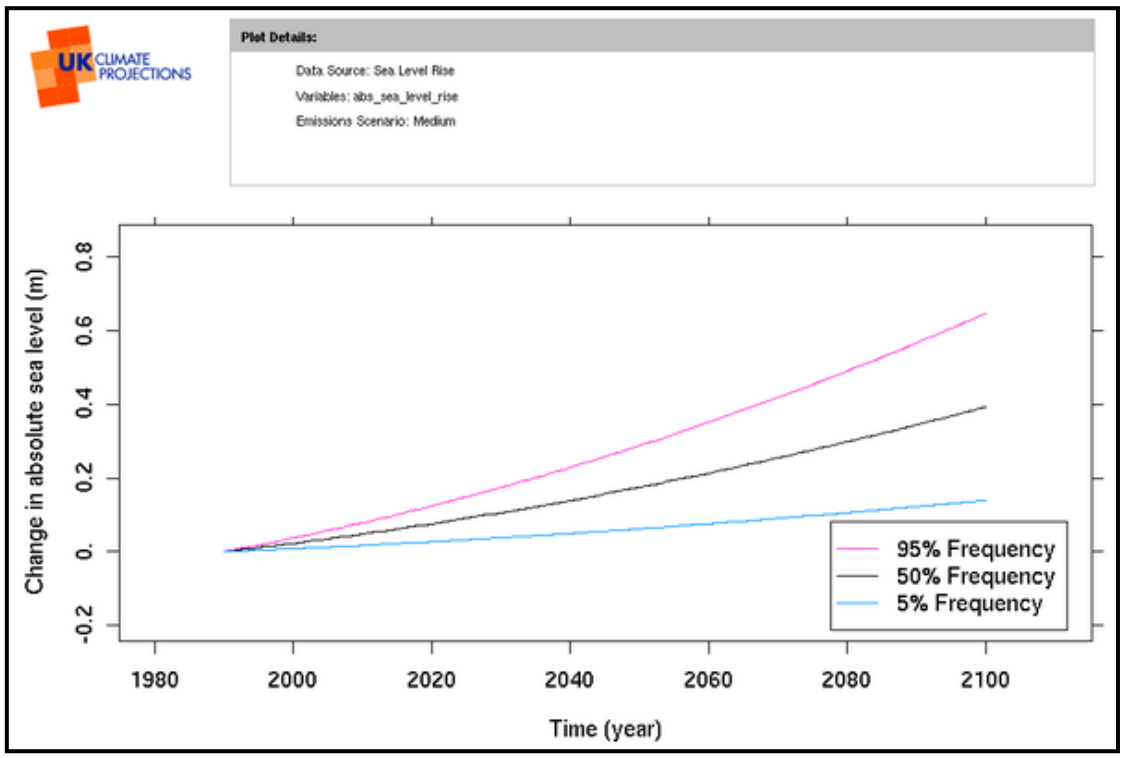


Figure 4. An image output with metadata above the plot.

5. Additional output files

5.1. Copyright file

All UKCP09 outputs are under Crown Copyright. This file contains the full copyright statement and is delivered in the zip file with all outputs.

6. Specific output types

This section explains some characteristics of the following specific output types:

- Weather Generator outputs
- Storm surge projections

6.1.1. Weather Generator outputs

Outputs produced directly from the UKCP09 Weather Generator (WG) do not undergo any post-processing by the UI. As a result the rest of this document does not apply to CSV and netCDF files from the WG.

Users should note the following characteristics are unique to the WG outputs:

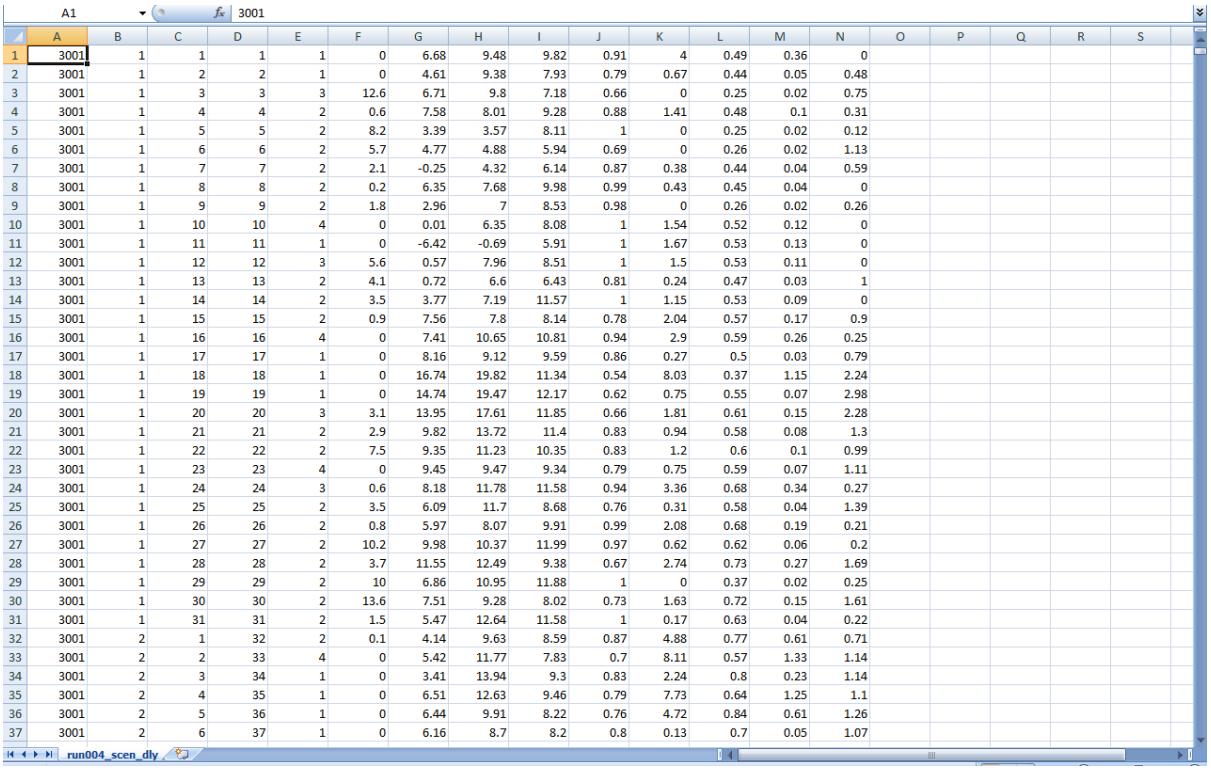
- A single file is written for every control and future climate run of the WG (of which there are a minimum of 100).
- WG output can total many GBs in size. Outputs for such large runs are grouped into multiple zip files. The reason for using multiple files is that it reduces the likelihood of timeout errors occurring when downloading very large files.
- CSV outputs from the WG do not include any header information or the first column in the main UKCP09 CSV files.
- An additional *column_headers.csv* file is delivered with WG CSV files. This describes the variable names for each column in all the CSV data files.

Please note that the variables in the WG outputs are different for *daily* and *hourly* modes of running.

Daily WG data includes the following variables:

Year: this is always given as starting at 3001 and runs through the number of years requested at a daily time step until the end year is reached.
Month: the month, values ranging from 1-12 (January-December).
Day: the day of the month.
Day_count: the cumulative total number of days in the year, values ranging from 1-366.
Transition: this defines what kind of day it is in the weather generator, values range from 1-4: 1 – indicates a dry day followed by a dry day; 2 - wet day followed by wet day; 3 – dry day followed by wet day; 4 – wet day followed by dry day.
Mean total daily precipitation rate (units: mm/day).
Minimum daily temperature, units: degrees C.
Maximum daily temperature, units: degrees C.
Vapour pressure, units: hPa.
Relative humidity, units: %, but these are given as decimal values in the file, so to get percentages these will need to be multiplied by 100.
Sunshine hours, units: hours (0-24).
Downward diffuse radiation, units: W/m^2.
Direct radiation, units: W/m^2.
Potential evapo-transpiration, units: mm/day.

Figure 5 shows an example WG CSV output viewed as a spreadsheet.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	3001	1	1	1	1	0	6.68	9.48	9.82	0.91	4	0.49	0.36	0
2	3001	1	2	2	1	0	4.61	9.38	7.93	0.79	0.67	0.44	0.05	0.48
3	3001	1	3	3	3	12.6	6.71	9.8	7.18	0.66	0	0.25	0.02	0.75
4	3001	1	4	4	2	0.6	7.58	8.01	9.28	0.88	1.41	0.48	0.1	0.31
5	3001	1	5	5	2	8.2	3.39	3.57	8.11	1	0	0.25	0.02	0.12
6	3001	1	6	6	2	5.7	4.77	4.88	5.94	0.69	0	0.26	0.02	1.13
7	3001	1	7	7	2	2.1	-0.25	4.32	6.14	0.87	0.38	0.44	0.04	0.59
8	3001	1	8	8	2	0.2	6.35	7.68	9.98	0.99	0.43	0.45	0.04	0
9	3001	1	9	9	2	1.8	2.96	7	8.53	0.98	0	0.26	0.02	0.26
10	3001	1	10	10	4	0	0.01	6.35	8.08	1	1.54	0.52	0.12	0
11	3001	1	11	11	1	0	-6.42	-0.69	5.91	1	1.67	0.53	0.13	0
12	3001	1	12	12	3	5.6	0.57	7.96	8.51	1	1.5	0.53	0.11	0
13	3001	1	13	13	2	4.1	0.72	6.6	6.43	0.81	0.24	0.47	0.03	1
14	3001	1	14	14	2	3.5	3.77	7.19	11.57	1	1.15	0.53	0.09	0
15	3001	1	15	15	2	0.9	7.56	7.8	8.14	0.78	2.04	0.57	0.17	0.9
16	3001	1	16	16	4	0	7.41	10.65	10.81	0.94	2.9	0.59	0.26	0.25
17	3001	1	17	17	1	0	8.16	9.12	9.59	0.86	0.27	0.5	0.03	0.79
18	3001	1	18	18	1	0	16.74	19.82	11.34	0.54	8.03	0.37	1.15	2.24
19	3001	1	19	19	1	0	14.74	19.47	12.17	0.62	0.75	0.55	0.07	2.98
20	3001	1	20	20	3	3.1	13.95	17.61	11.85	0.66	1.81	0.61	0.15	2.28
21	3001	1	21	21	2	2.9	9.82	13.72	11.4	0.83	0.94	0.58	0.08	1.3
22	3001	1	22	22	2	7.5	9.35	11.23	10.35	0.83	1.2	0.6	0.1	0.99
23	3001	1	23	23	4	0	9.45	9.47	9.34	0.79	0.75	0.59	0.07	1.11
24	3001	1	24	24	3	0.6	8.18	11.78	11.58	0.94	3.36	0.68	0.34	0.27
25	3001	1	25	25	2	3.5	6.09	11.7	8.68	0.76	0.31	0.58	0.04	1.39
26	3001	1	26	26	2	0.8	5.97	8.07	9.91	0.99	2.08	0.68	0.19	0.21
27	3001	1	27	27	2	10.2	9.98	10.37	11.99	0.97	0.62	0.62	0.06	0.2
28	3001	1	28	28	2	3.7	11.55	12.49	9.38	0.67	2.74	0.73	0.27	1.69
29	3001	1	29	29	2	10	6.86	10.95	11.88	1	0	0.37	0.02	0.25
30	3001	1	30	30	2	13.6	7.51	9.28	8.02	0.73	1.63	0.72	0.15	1.61
31	3001	1	31	31	2	1.5	5.47	12.64	11.58	1	0.17	0.63	0.04	0.22
32	3001	2	1	32	2	0.1	4.14	9.63	8.59	0.87	4.88	0.77	0.61	0.71
33	3001	2	2	33	4	0	5.42	11.77	7.83	0.7	8.11	0.57	1.33	1.14
34	3001	2	3	34	1	0	3.41	13.94	9.3	0.83	2.24	0.8	0.23	1.14
35	3001	2	4	35	1	0	6.51	12.63	9.46	0.79	7.73	0.64	1.25	1.1
36	3001	2	5	36	1	0	6.44	9.91	8.22	0.76	4.72	0.84	0.61	1.26
37	3001	2	6	37	1	0	6.16	8.7	8.2	0.8	0.13	0.7	0.05	1.07

Figure 5. Example of a daily CSV output from the Weather Generator,

Hourly WG data includes the following variables:

Year: this is always given as starting at 3001 and runs through the number of years requested at a hourly time step until the end year is reached.
Month: the month, values ranging from 1-12 (January-December).
Day: the day of the month.
Hour: the hour of the day.
Mean total hourly precipitation rate (units: mm/day).
Mean hourly temperature, units: degrees C.
Vapour pressure, units: hPa.
Relative humidity, units: %, but these are given as decimal values in the file, so to get percentages these will need to be multiplied by 100.
Sunshine hours, units: hours (0-24).
Downward diffuse radiation, units: W/m^2.
Direct radiation, units: W/m^2.

6.1.2. Storm surge projections

The storm surge projections are provided for a single location (grid box). The main variable is *skew surge* which is defined at 4 return levels (2, 10, 20 and 50 years) against *uncertainly level* (5%, 50% and 95%).

The dataset also contains the *statistical significance* of the skew surge projections.

In netCDF outputs the storm surge projections this additional variable is written as standard netCDF variables.

7. Grids and Spatial averages used in UKCP09

When building requests in the UI you will be required make spatial selections in various places. There are two types of location (called *spatial average* in the UI):

- Gridded data (such as the MOHC 25km rotated pole grid or the sea level rise grid)
- Aggregated areas (such as the river basins or marine regions)

7.1. Further spatial information

For details of the grids that UKCP09 outputs are defined against please see the grids documentation pages at:

http://ukclimateprojections-ui.metoffice.gov.uk/ui/docs/grids/

7.2. Handling the MOHC rotated grid in UKCP09 outputs

The MOHC 25km rotated grid has a rotated North Pole location. The grid therefore looks as though it has all been rotated slightly in an anti-clockwise direction (see Figure 6). This section explains how the grid is defined in the UKCP09 CSV and CF-netCDF files.

Figure 6. The MOHC 25km rotated grid

Since the grid is rotated it is defined on dimensions called *rotated latitude* and *rotated longitude*. The real latitudes and longitudes are unique for every single grid point because the grid boxes are not aligned with regular latitudes and longitudes.

There are two types of request that use the rotated grid in UKCP09. The first type is when a single grid box is requested. In this case details of the rotated grid are not particular useful. The grid box is identified by a grid box ID number. The rotated is provided on 52 y-values and 39 x-values. There are 2028 values in total with 440 representing valid land points. The grid box ID of any particular point can be calculated from the formula:

```
grid_box_id = (y_index * nx) + x_index
```

Where `nx` is the number of x-values (39 in this case).

The second type of request is for a mapped output of a climate variable at a given probability level. These outputs are defined over the entire UK region or a rectangular

subset of it. Depending on the file format the spatial information is represented in different ways.

7.2.1. Rotated grid data in CSV

The UKCP09 CSV format represents the 2D array of values in a table as discussed above. The CSV header provides some information about the dimensions (as documented in chapter 3. However, the actual values of the rotated latitude and longitudes are replaced with simple index values in the CSV files. This was decided in consultation with the UKCP09 Users' Panel to avoid the possibility that users would misinterpret rotated coordinates as real world coordinates. In order to look up the real values users should consult the 25km grid documentation at:

http://ukclimateprojections-ui.metoffice.gov.uk/ui/docs/grids/prob_land_25km_rotated

7.2.2. Rotated grid data in CF-netCDF

When described in CF-netCDF the rotated grid follows the recommendations of the CF Conventions as follows:

- The climate variable is defined against *rotated latitude* and *rotated longitude*
- The climate variable includes a *coordinates* attribute that includes the ids of the additional coordinate variables *real latitude* and *real longitude*.
- The *real latitude* and *real longitude* coordinate variables are defined against the *rotated latitudes* and *rotated longitudes.*
- The climate variable includes a *grid_mapping* attribute set to the name of the rotated pole variable in the file
- The rotated pole variable is defined with the modified location of the Pole.

Further details are available on the CF web site at:

http://cf-pcmdi.llnl.gov/documents/cf-conventions/1.4/ch05s06.html

8. References

NASA Ames format web site at BADC:
http://badc.nerc.ac.uk/help/formats/NASA-Ames

NASA Ames Definition Document:
http://badc.nerc.ac.uk/help/formats/NASA-Ames/G-and-H-June-1998.html

NetCDF Home Page (Unidata):
http://www.unidata.ucar.edu/software/netcdf/

Climate and Forecasts (CF) Conventions:
http://cf-pcmdi.llnl.gov/

Software for NetCDF page:
http://www.unidata.ucar.edu/software/netcdf/

BADC page on CF-NetCDF:
http://badc.nerc.ac.uk/help/formats/netcdf/

Appendices

Appendix 1: Acronyms used in this document

UI	User Interface
MOHC	Met Office Hadley Centre
NetCDF-CF	Network Common Data Format - Climate and Forecasts Metadata Conventions
BADC	British Atmospheric Data Centre
TTOM	Task Team on Outputs and Metadata
UKCIP	UK Climate Impacts Programme
UKCP09	UK Climate Projections
GIS	Geographical Information Systems
NASA Ames	NASA Ames file format
CSV	Comma-separated variables
ESRI	Environmental Systems Research Institute

Appendix 2: Reducing data structure complexity

All UKCP09 datasets are stored on our servers in netCDF files. The stored data structures are different from how they are delivered by the UI. When a request for data comes into the UI the data is read from the files and then re-processed according to a set of rules. These rules typically reduce the complexity of the data by converting it to a 1-dimensional or 2-dimensional structure. This is done by two methods:

1. Combining variables with dimensions to reduce the number of dimensions
2. Flattening (squeezing) any dimensions that only contain one value

1. Combining variables and dimensions together (CSV only)

Many variables are defined against a number of dimensions. For example temperature can be defined against (emissions scenario, temporal average, time period). This 3D data structure would be hard to represent in a table. We have taken the approach of combining variables and dimensions in to new variables in order to reduce the number of dimensions.

The example below demonstrates the principle with a simple 1-dimensional variable:

- temperature(jan, feb) would become temperature_in_jan, and temperature_in_feb

The UI software analyses each 2D output type and wherever possible pre-process the data to combine variables with one axis.

2. Flattening dimensions that only contain one value (CSV and netCDF)

In some cases there are a number of dimensions that only contain one value, such as the high emissions scenario, 2070-2099 time period and temporal average of July. When this is the case the metadata in the output files retains these parameters so these “singleton” dimensions can be flattened out of the variable to simplify it.

This is also relevant to netCDF because many software tools can only handle variables defined against a limited number of dimensions. Since some of the UKCP09 datasets are 6-dimensional then dimensions that only contain one value (singleton dimensions) are removed and are referenced by the coordinates attribute to retain a linkage.

The table below lists the various output types by data source, their dimensionality, and the appropriate UKCP09 sub-format used to render them.

Codes used in table:

TA	Temporal Average
Sample	Model Variant
TP	Time Period
ES	Emissions Scenario
Loc	Location
Rlat	Rotated grid latitudes
Rlon	Rotated grid longitudes
PC	Percentile

My code	Data Source	Subset details	Number of variables	Dimensions and lengths	Number of dimensions	Flattened dimensions	Axis combined with variable to make easy to tabulate	Sub-format	Comments	Non-standard variables, attributes and their rules
1.1	Prob land, region/river	Sampled data	1-8	TA = 1-17 Sample index = 1-10000	2 (flattened into 1)	Loc, ES, TP	TA	1001	We are combining variable and TA into variable names so it follows the original UKCP09 dummy data format. Each variable/TA combination will appear in a column in the CSV file.	sample_index axis needs overwriting with sample_id coordinate variable.
1.2	Prob land, 25km grid box	Sampled data	1-8	TA = 1-17 Sample index = 1-10000	2 (flattened into 1)	Loc, ES, TP	TA	1001	We are combining variable and TA into variable names so it follows the original UKCP09 dummy data format. Each variable/TA combination will appear in a column in the CSV file. NOTE: Reference file will be available for grid box real lat/lon lookups.	sample_index axis needs overwriting with sample_id coordinate variable. Add a comment about grid_box_id and where it can be looked up.
1.3	Prob land for joint prob plot	Sampled data	2	Sample = 10000	1	Loc, ES, TP, TA		1001	Data must just be two vectors (one for each variable).	sample axis needs overwriting with sample_id coordinate variable
1.4	Prob land, single	PDF/CDF/Raw data, 1-	1	ES = 1-3 PC = 131	1-2 (flattened into 1)	Loc, TP, TA, [ES]	ES	1001	Combine Var and Em scen so it is possible to represent the 2D	Add a comment (for 25km grid box) about

	location	3 ES							data as 1D, the same as one em scen version.	grid_box_id and where it can be looked up.
1.5	Prob land, for plume plot	Plume/Raw data	1	TP = 7 PC = 8	2 (flattened into 1)	Loc, ES, TA	TP	1001	Combine Var and Time Period into variables.	Add a comment (for 25km grid box) about grid_box_id and where it can be looked up.
1.6	Prob land, 25km grid bounding box	Map of single percentile over bounding box/Raw data	1	Rlat = 2-52 Rlon = 2-39	2	ES, TA, PC, TP (flattened but retained as axis)		3010	Best format for representing gridded data. Note that Time Period axis is maintained so that data can be written to 3010.	Add a comment (for 25km grid box) about grid_box_id and where it can be looked up.
2.1	Prob marine, marine region	As 1.1	As 1.1	As 1.1	As 1.1	As 1.1	As 1.1	As 1.1	As 1.1	As 1.1
2.2	Prob marine for joint prob plot	As 1.3	As 1.3	As 1.3	As 1.3	As 1.3	As 1.3	As 1.3	As 1.3	As 1.3
2.3	Prob marine, single location	As 1.4	As 1.4	As 1.4	As 1.4	As 1.4	As 1.4	As 1.4	As 1.4	As 1.4
2.4	Prob marine, for plume plot	As 1.5	As 1.5	As 1.5	As 1.5	As 1.5	As 1.5	As 1.5	As 1.5	As 1.5
2.5	Prob marine, for all UK	1.7	As 1.6	Loc = 1-	As 1.6	As 1.6	As 1.6	As 1.6	As 1.6	As 1.6
3.1	Wxgen output	daily	9	Sample = 100-1000 Time = 10000ish	2 (but 1 – see comment)	Loc, ES, TP, TA		Like 1001 – as raw WG format	Each WG output is written to a separate file (i.e. for each Sample) so the 2D data structure will actually be rendered into a set of 1D output files.	as is
3.2	Wxgen output	hourly	7	Sample = 100 Time = 100000ish	2 (but 1 – see comment)	Loc, ES, TP, TA		Like 1001 – as raw WG format	Each WG output is written to a separate file (i.e. for each Sample) so the 2D data structure will actually be rendered into a set of 1D output files.	as is
5.1	Storm surge trend plot	Trend plot/Raw data	1 (plus flag)	Ret Level = 4 Percentile = 3	2 (flattened into one)	Loc	Ret Level	1001	Combine the return level with the variable name to make it 4 variables that are 1D.	H++ is read and put in comments. Significance is in comments.

6.1	Sea Level Rise plume plot	Absolute SLR, Plume/Raw data	1	TP = 111 PC = 3	2	ES	PC	1001	Percentile will be combined with variable as there are only 3.	
6.2	Sea Level Rise plume plot	Relative SLR, Plume/Raw data	1	TP = 111 PC = 3	2 (flattened into one)	ES, Loc	PC	1001		