Climate Predictability Tool (CPT) User Guide
# Table of Contents

1. Introduction .............................................................................................................. 1  
2. System Requirements .............................................................................................. 2  
I. Input Data ................................................................................................................... 3  
   3. Input Data Format .................................................................................................... 5  
      V10 File Formats .................................................................................................... 5  
      Common fields ......................................................................................................... 5  
      Gridded Data Format .............................................................................................. 6  
      Station Data Format ............................................................................................... 9  
      Unreferenced Data Format .................................................................................... 10  
      Older File Formats (File Format before version 10) .............................................. 11  
4. Downloading CPT Data From IRI Data Library ..................................................... 13  
   Downloading Sea Surface Temperature Data .......................................................... 13  
   Downloading ECHAM4.5 Forecast Data ................................................................. 17  
   Download Dual Field Station Data ........................................................................... 19  
II. CPT Windows Version Tutorial ............................................................................... 22  
   5. Getting Started ....................................................................................................... 24  
   6. Selecting Analysis Method ..................................................................................... 25  
      Choose the analysis to perform: PCR, CCA or MLR ............................................. 25  
   7. Select Input Datasets .............................................................................................. 26  
      Select Input Datasets ............................................................................................ 26  
   8. Set Program Parameters ....................................................................................... 27  
      Specify X domain and EOF options ...................................................................... 27  
      Specifying Y Domain and EOF options ............................................................... 27  
      Setting the Training Period, Cross-Validation Window ........................................ 27  
      Setting Missing Values ......................................................................................... 28  
      Set Other Options .................................................................................................. 28  
      Save Program Parameters ..................................................................................... 29  
9. Performing the Analysis ........................................................................................... 30  
    Analysis Progress ..................................................................................................... 30  
10. CPT Results ............................................................................................................. 31  
    CPT Graphics .......................................................................................................... 31  
    Saving the Graphics ................................................................................................. 31  
    Customising the graphics ....................................................................................... 32  
    Saving Results to Files ............................................................................................. 32  
11. CPT Results Analysis ............................................................................................ 34  
    ................................................................................................................................. 34  
12. CPT Forecasts .......................................................................................................... 35  
    Selecting Forecast File ............................................................................................. 35  
    Setting Forecast Options ......................................................................................... 35  
    Forecast Results ....................................................................................................... 36  
      Forecast Results: Series ....................................................................................... 37  
      Forecast Results: Maps ......................................................................................... 37  
      Forecast Results: probabilities of exceedance .................................................... 37  
    Changing Forecast Settings .................................................................................... 38  
III. Specific CPT Menu Items ......................................................................................... 43  
13. File ........................................................................................................................... 46  
    New ......................................................................................................................... 46  
    Open ......................................................................................................................... 46  
    Save ......................................................................................................................... 46  
    Save As ..................................................................................................................... 46  
    Output Results ........................................................................................................ 46
Climate Predictability
Tool (CPT) User Guide

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exit</td>
<td>47</td>
</tr>
<tr>
<td>Close</td>
<td>47</td>
</tr>
<tr>
<td>14. Edit</td>
<td>48</td>
</tr>
<tr>
<td>X Data Domain</td>
<td>48</td>
</tr>
<tr>
<td>Y Data Domain</td>
<td>48</td>
</tr>
<tr>
<td>15. Actions</td>
<td>49</td>
</tr>
<tr>
<td>Calculate</td>
<td>49</td>
</tr>
<tr>
<td>Cross-Validated</td>
<td>49</td>
</tr>
<tr>
<td>Retroactive</td>
<td>49</td>
</tr>
<tr>
<td>Reset</td>
<td>50</td>
</tr>
<tr>
<td>16. Tools</td>
<td>51</td>
</tr>
<tr>
<td>Validation</td>
<td>51</td>
</tr>
<tr>
<td>Cross-Validated</td>
<td>51</td>
</tr>
<tr>
<td>Retroactive</td>
<td>52</td>
</tr>
<tr>
<td>Verification</td>
<td>52</td>
</tr>
<tr>
<td>Contingency Tables</td>
<td>54</td>
</tr>
<tr>
<td>Modes</td>
<td>54</td>
</tr>
<tr>
<td>Forecasts</td>
<td>56</td>
</tr>
<tr>
<td>17. Options</td>
<td>60</td>
</tr>
<tr>
<td>X EOF Options</td>
<td>60</td>
</tr>
<tr>
<td>Y EOF Options</td>
<td>60</td>
</tr>
<tr>
<td>CCA Options</td>
<td>60</td>
</tr>
<tr>
<td>Climatological Period</td>
<td>60</td>
</tr>
<tr>
<td>Tailoring</td>
<td>60</td>
</tr>
<tr>
<td>Data</td>
<td>61</td>
</tr>
<tr>
<td>Transform Y Data</td>
<td>61</td>
</tr>
<tr>
<td>Zero-Bound</td>
<td>61</td>
</tr>
<tr>
<td>Missing Values</td>
<td>62</td>
</tr>
<tr>
<td>Resampling Settings</td>
<td>62</td>
</tr>
<tr>
<td>Forecast Settings</td>
<td>62</td>
</tr>
<tr>
<td>Graphics</td>
<td>62</td>
</tr>
<tr>
<td>Crosses on Graphs</td>
<td>62</td>
</tr>
<tr>
<td>Mask Land</td>
<td>62</td>
</tr>
<tr>
<td>Mask Lakes</td>
<td>63</td>
</tr>
<tr>
<td>Black and White</td>
<td>63</td>
</tr>
<tr>
<td>Reverse Colours</td>
<td>63</td>
</tr>
<tr>
<td>Vertical Lines on Graph</td>
<td>63</td>
</tr>
<tr>
<td>18. View</td>
<td>64</td>
</tr>
<tr>
<td>Canonical Correlation Analysis (CCA)</td>
<td>64</td>
</tr>
<tr>
<td>Principal Components Regression (PCR)</td>
<td>64</td>
</tr>
<tr>
<td>Multiple Linear Regression (MLR)</td>
<td>64</td>
</tr>
<tr>
<td>19. Customize</td>
<td>65</td>
</tr>
<tr>
<td>Scree Plot</td>
<td>65</td>
</tr>
<tr>
<td>Cumulative</td>
<td>65</td>
</tr>
<tr>
<td>Logarithmic Axis</td>
<td>65</td>
</tr>
<tr>
<td>Broken Stick</td>
<td>65</td>
</tr>
<tr>
<td>Title</td>
<td>65</td>
</tr>
<tr>
<td>IV. CPT Linux Version Tutorial</td>
<td>66</td>
</tr>
<tr>
<td>20. Linux Version Introduction</td>
<td>68</td>
</tr>
<tr>
<td>21. Linux Version Installation</td>
<td>69</td>
</tr>
<tr>
<td>System and Compiler Requirements</td>
<td>69</td>
</tr>
<tr>
<td>How to Build CPT</td>
<td>69</td>
</tr>
<tr>
<td>How to Run CPT</td>
<td>69</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>lxxii</td>
</tr>
</tbody>
</table>
List of Figures

4.1. ERSST Data Selection ................................................................. 14
4.2. Data Selection .................................................................................. 15
4.3. Range Selection ................................................................................ 16
4.4. Save CPT Data File ........................................................................... 17
4.5. ECHAM Expert Mode ...................................................................... 18
4.6. ECHAM Expert Mode Data File ....................................................... 19
4.7. Dual Fields Station Data ................................................................... 20
4.8. Dual Fields Station Data Info ........................................................... 21
6.1. Main Menu ......................................................................................... 25
7.1. Input Data ........................................................................................... 26
8.1. X Domain Selection ........................................................................... 27
8.2. X EOF Options .................................................................................. 27
8.3. CCA Settings ..................................................................................... 28
8.4. Missing Values .................................................................................. 28
8.5. Options ............................................................................................... 29
21.1. PCR ................................................................................................. 71
Chapter 1. Introduction

The Climate Predictability Tool (CPT) provides a Windows package for constructing a seasonal climate forecast model, performing model validation, and producing forecasts given updated data. Its design has been tailored for producing seasonal climate forecasts using model output statistic (MOS) corrections to climate predictions from general circulation model (GCM), or for producing forecasts using fields of sea-surface temperatures. Although the software is specifically tailored for these applications, it can be used in more general settings to perform canonical correlation analysis (CCA) or principal components regression (PCR) on any data, and for any application. Comments and requests for changes and developments can be emailed to <cpt@iri.columbia.edu>.


Download the Appendix: AppendixA.pdf
Chapter 2. System Requirements


**Minimum diskspace:** Less than 10 Mb for program files

**Recommended minimum screen resolution:** 1024 x 768

Memory is allocated dynamically and depends on the size of the datasets. The software does not impose its own memory limits, although it is currently unable to read in more than 36156 or output more than 1128 longitudes / stations / series, and occasional memory allocation problems may cause the program to crash if large datasets are used. The program will work at a screen resolution of 800 x 600, but some of the graphics may be difficult to view. The program does not work on 64-bit processors. There have been a few compatibility problems with some machines running Windows Vista and Windows 7; we are working to resolve these.

A source code version of the software, written in Fortran90 and C, is available on request to <cpt@iri.columbia.edu>, which should be compilable on most platforms, and which can be used for batch work. See the Chapter 20, *Linux Version Introduction* CPT Linux Version Tutorial.
Part I. Input Data
# Table of Contents

3. Input Data Format ............................................................................................... 5  
   V10 File Formats ............................................................................................... 5  
   Common fields .................................................................................................... 5  
   Gridded Data Format ......................................................................................... 6  
   Station Data Format ........................................................................................... 9  
   Unreferenced Data Format ................................................................................ 10  
   Older File Formats (File Format before version 10) .............................................. 11  
4. Downloading CPT Data From IRI Data Library .................................................. 13  
   Downloading Sea Surface Temperature Data ..................................................... 13  
   Downloading ECHAM4.5 Forecast Data ............................................................ 17  
   Download Dual Field Station Data ...................................................................... 19
Chapter 3. Input Data Format

V10 File Formats

Common fields

All three formats (gridded, station and unreferenced data formats) must begin with the following lines.

The **first line** of an input file is always:

```xml
xmlns:cpt=http://iri.columbia.edu/CPT/v10/
```

This tag defines an XML namespace for the "cpt" prefix to be used in subsequent lines. This line should be copied exactly as is at the top of all CPT version 10 input files.

The **second line** indicates the number of "fields" in the file:

```xml
cpt:nfields=1
```

A "field" is a set of variables for the same meteorological parameter measured at different locations; for example, rainfall measurements at a set of stations, or a grid of sea-surface temperature records. In older versions of CPT only one field was permitted, but in CPT version 10 and subsequent versions, multiple fields can be used. How the different fields are set out (AEC: not sure what this means) in the file depends on the file type, as described in the following sections, but for unreferenced files, only one field is permitted.

The **third line** starts with the CPT tag

```xml
```

where "T" represents time, followed by a list of all the dates for which data are available in the file. The format of the date follows ISO8601 format standards. The year is listed first, followed (optionally) by the month, and then the day (again optionally). The year month and day are separated by dashes.

For example, 30 September 2009 would be indicated as 2009-09-30.

If the day is missing, it is implied that the data are representative of the entire month (either as an average or cumulative total). For example, if the file contains monthly rainfall data for September, the days would be omitted in the date format, and September 2009 would be indicated as 2009-09.

If only the year is listed (omitting both the month and day), CPT interprets the data as representative of the whole year. (Note: In older versions, CPT assumes the month is undefined when only the year is listed.)

If the data represent seasonal averages, for example, it is possible to indicate the season in the date format by using a "/" separator for the start and end dates. For example, a July - September 2009 3-month average would be represented by 2009-07/2009-09. However, since the year is the same in both cases, this can be abbreviated to 2009-07/09. It is implicit that the data are from 01 July 2009 to 30 September 2009, and so the days of the month are not required.

If the three-month average spans the year-end, [e.g. December to February (DJF)] then the year will need to be included for the beginning and the end dates. For example, 2008-12/2009-02 is correct, but 2008-12/02 is invalid. When CPT requests a start date, the beginning of the season should be indicated for seasonal averages. So, for example, to start in DJF 1971/72, the start date is December 1971.

The **fourth line** contains a series of CPT tags that set information about the block of data immediately following. This information depends on the file structure, and so this line is described separately for each
format. The tags can appear in any order, but some tags are compulsory. Each tag is preceded by "cpt" followed by a colon and the name of the tag, then "=" and the value that the tag takes. Tags are also case sensitive.

**Gridded Data Format**

**New Gridded Data Format**

Gridded data files organise data in blocks with each block representing a common period and field, with the rows representing latitudes and the columns longitudes.

The fourth line contains the following compulsory tags.

* cpt:T=time period for which the data are valid
  (in the ISO8601 format described above)
* cpt:nrow=number of latitudes
* cpt:ncol=number of longitudes
* cpt:row=Y (representing latitudes)
* cpt:col=X (representing longitudes)

Optional tags are:

* cpt:field=abbreviated name of the field variable
* cpt:S=start date (for GCM outputs indicating when the forecast was started (the compulsory cpt:T tag gives the target season of the forecast)
* cpt:units=units in which the data are stored
* cpt:missing=missing value flag

Any additional tags are ignored.

Immediately after this tag line (fifth line) there is a line listing all the longitudes from west to east. The longitudes must be between -180 and 360, with negative values representing the western hemisphere. There must be ncol longitudes.

The data follow in the subsequent nrow rows, with the first value representing the latitude. Latitudes must be between 90 and -90, with negative values representing the southern hemisphere. The data can be from north to south or from south to north. If there is only one field and there are no EOF extensions (e.g. more than one season per year) then the data for the subsequent date should follow. All tag values should remain the same except for cpt:T=time. For example, if the data are SST anomalies in Celsius, the file might look something like the following:

```xml
 -18.75 -16.25 -13.75 -11.25
 -58.75 -2.40977 -1.40228 -1.01278 -1.03197
 -61.25 -3.65581 -3.18779 -2.28418 -1.45435
 -63.75 -3.54459 -3.04635 -2.07969 -1.32451
 -18.75 -16.25 -13.75 -11.25
 -58.75 -2.22299 -0.68736 0.27909 -0.42638
 -61.25 -1.13313 -0.56869 -1.02204 -1.09111
 -63.75 -1.40550 -1.10838 -0.37565 0.88614
```
If there is an EOF extension the additional dates must appear in the third line. For example, in the example below data for January and February of each year are included. Since CPT is looking for one value per year, the February data will be treated as an EOF extension. In effect the February data are an extra field, but since they have a different date rather than being a different meteorological parameter they are handled as an EOF extension. Although there are 4 dates listed, because the February values are recognised as EOF extensions, there are only two years of data available for the training period. CPT will identify the number of years in the file, and recognise that there is one EOF extension.

A file with **1 fields and one EOF extension** looks like as follows:

```xml
<xmlns:cpt=http://iri.columbia.edu/CPT/v10/
cpt:nfields=1
    -18.75 -16.25 -13.75 -11.25
    -58.75 -2.40977 -1.40228 -1.01278 -1.03197
    -61.25 -3.65581 -3.18779 -2.28418 -1.45435
    -63.75 -3.54459 -3.04635 -2.07969 -1.32451
    -18.75 -16.25 -13.75 -11.25
    -58.75 -1.34563 -0.23456 -0.98857 -0.37327
    -61.25 -2.23454 -3.21124 -1.99463 -2.44437
    -63.75 -4.00043 -3.05441 -2.12645 -1.40078
    -18.75 -16.25 -13.75 -11.25
    -58.75 -2.22299 -0.68736 0.27909 -0.42638
    -61.25 -1.13313 -0.56869 -1.02204 -1.09111
    -63.75 -1.40550 -1.10838 -0.37565 0.88614
    -18.75 -16.25 -13.75 -11.25
    -58.75 -2.34256 -0.23346 0.26537 -0.67278
    -61.25 -1.07636 -0.96765 -0.85677 -1.14531
    -63.75 -1.34254 -1.04553 -0.25831 0.82646
```

If there is second field the first data for the second field must come immediately after the first data for the first field, and must have the same date. They should come before the first data for any EOF extension. Thus the blocks of data should alternate between fields. Currently the latitudes and longitudes for the two fields must be identical, although there are plans to drop this restriction at a later date. The only tags that may change will therefore be `cpt:field=`, `cpt:units=`. Currently the missing value flag has to be the same for both fields. This restriction may change if there is sufficient demand.

Thus a file with **two fields** may look something like the following:

```xml
<xmlns:cpt=http://iri.columbia.edu/CPT/v10/
cpt:nfields=2

cpt:T 1979-01 1980-01

    -18.75 -16.25 -13.75 -11.25
    -58.75 -2.40977 -1.40228 -1.01278 -1.03197
    -61.25 -3.65581 -3.18779 -2.28418 -1.45435
    -63.75 -3.54459 -3.04635 -2.07969 -1.32451
    -18.75 -16.25 -13.75 -11.25
    -58.75 -2.22299 -0.68736 0.27909 -0.42638
    -61.25 -1.13313 -0.56869 -1.02204 -1.09111
    -63.75 -1.40550 -1.10838 -0.37565 0.88614
```

Thus a file with **two fields** may look something like the following:
Input Data Format

-63.75 4.45351 2.67322 2.56088 3.67663
-58.75 -2.22299 -0.68736 0.27909 -0.42638
-61.25 -1.13313 -0.56869 -1.02204 -1.09111
-63.75 -1.40550 -1.10838 -0.37565 0.88614
-18.75 -16.25 -13.75 -11.25
-58.75 1.98867 0.65446 0.54686 0.54322
-61.25 3.77636 1.56422 0.76835 1.16771
-63.75 1.33455 1.65456 0.79965 1.23142

A file with **2 fields and one EOF extension** must be ordered as follows:

```
xmns:cpt=http://iri.columbia.edu/CPT/v10/
cpt:nfields=2
-58.75 -2.40977 -1.40228 -1.01278 -1.03197
-61.25 -3.65581 -3.18779 -2.28418 -1.45435
-63.75 -3.54459 -3.04635 -2.07969 -1.32451
-18.75 -16.25 -13.75 -11.25
-58.75 1.23456 0.65246 0.98578 0.67533
-61.25 2.86344 3.26537 1.04836 2.76536
-63.75 4.45351 2.67322 2.56088 3.67663
-58.75 -1.34563 -0.23456 -0.98857 -0.37327
-61.25 -2.23454 -3.21124 -1.99463 -2.44437
-63.75 -4.00043 -3.05441 -2.12645 -1.40078
-18.75 -16.25 -13.75 -11.25
-58.75 1.09654 0.76382 0.87644 0.87644
-61.25 2.02675 3.93176 1.65738 2.65437
-63.75 4.47507 2.65490 2.76113 3.65425
-58.75 -2.22299 -0.68736 0.27909 -0.42638
-61.25 -1.13313 -0.56869 -1.02204 -1.09111
-63.75 -1.40550 -1.10838 -0.37565 0.88614
-18.75 -16.25 -13.75 -11.25
-58.75 1.88867 0.65446 0.54686 0.54322
-61.25 3.77636 1.56422 0.76835 1.16771
-63.75 1.33455 1.65456 0.79965 1.23142
-63.75 -3.54459 -3.04635 -2.07969 -1.32451
```
**Input Data Format**

<table>
<thead>
<tr>
<th>-58.75</th>
<th>1.76456</th>
<th>0.65417</th>
<th>0.13454</th>
<th>0.46289</th>
</tr>
</thead>
<tbody>
<tr>
<td>-61.25</td>
<td>3.76567</td>
<td>1.86356</td>
<td>0.54366</td>
<td>1.97641</td>
</tr>
<tr>
<td>-63.75</td>
<td>1.52255</td>
<td>1.54326</td>
<td>0.65477</td>
<td>1.76533</td>
</tr>
</tbody>
</table>

**Station Data Format**

Station data files organise data in blocks with each block representing a field, with the rows representing time, and the columns stations. The tags are very similar to those for gridded data. Compulsory tags are:

- *cpt:T=time period for which the data are valid (in the ISO8601 format described above)*
- *cpt:nrow=number of time steps*
- *cpt:ncol=number of stations*
- *cpt:row=T (representing time)*
- *cpt:col=station*

The number of time steps is usually the number of years of data available, but if there are EOF extensions, the number of time steps will increase. For example, if data are available for January and February 1971 to 2000 there are two months (one EOF extension) and thirty years of data, making a total of 60 time steps.

Optional tags are:

- *cpt:field=abbreviated name of the field variable*
- *cpt:units=units in which the data are stored*
- *cpt:missing=missing value flag*

Any additional tags are ignored. Immediately beneath the tag-line should be a line containing the names, or some kind of unique identifying symbol, for each of the ncol stations. The name for each station must not be longer than 16 characters, and should contain no spaces. The latitudes and longitudes of each station then follow on the next two lines. The line for the longitudes must begin with "cpt:X" and that for the latitudes with "cpt:Y". It does not matter whether the latitudes or the longitudes are listed first. As with the gridded data, latitudes must be between 90 and -90 with negative latitudes representing the southern hemisphere, and longitudes must be between -180 and 360, with negative longitudes representing the western hemisphere. Additional tagged lines may follow, as in the example below, which includes a line for elevation and one for province. These lines are currently ignored, and are not included in any output files. The data are arranged in columns by station as the "cpt:col" tag indicates, but the first column must contain the corresponding date. The date format is the same as for gridded data. The seasonal average format with the "/" separator is permissible.

```
xmns:cpt=http://iri.columbia.edu/CPT/v10/ cpt:nfields=1
A       B       C       D       E       F
| 122.006 | 121.050 | 121.633 | 120.600 | 121.557 |
| 5 m     | 10 m    | 1503 m  | 25 m    | 6 m     |
```

2000-01 : -2.0   121.5  124.0    90.0    95.5
2000-02 : 345.5  87.5   224.0   173.5   34.5
2001-01 : 428.5  93.5   301.0   104.0   -2.0
2001-02 : 341.0  126.0  156.0   130.0   62.5
2002-01 : -2.0   34.0   185.5   75.5    71.5
2002-02 : 117.5  72.0   94.0    216.0   67.5
2003-01 : 143.5  34.5   241.0   134.5   4.5
Input Data Format

Any additional fields are included immediately after the data, starting with a new tag-line. All the tags must be exactly the same as for the first field except for "cpt:field" and "cpt:units". An example is shown below.

```
xmns:cpt=http://iri.columbia.edu/CPT/v10/  
cpt:nfields=2


cpt:field=prcp, cpt:nrow=8, cpt:ncol=6, cpt:row=T, cpt:col=station, cpt:units=mm/month

A B C D E
A 122.006 121.050 121.633 120.600 121.557
C 5 m 10 m 1503 m 25 m 6 m
Province Z Y X W V
2000-01 -2.0 121.5 124.0 98.0 95.5
2000-02 345.5 87.5 224.0 173.5 34.5
2001-01 428.5 93.5 301.0 104.0 -2.0
2001-02 341.0 126.0 156.0 130.0 62.5
2002-01 -2.0 34.0 185.5 75.5 71.5
2002-02 117.5 72.0 94.0 216.0 67.5
2003-01 143.5 34.5 241.0 134.5 4.5
2003-02 355.0 81.5 137.5 62.5 7.0
```


```
A B C D E
A 122.006 121.050 121.633 120.600 121.557
C 5 m 10 m 1503 m 25 m 6 m
Province Z Y X W V
2000-01 -2.0 23.3 14.4 22.1 21.3
2000-02 20.4 24.9 17.6 23.6 23.6
2001-01 23.3 24.8 15.8 25.4 -2.0
2001-02 25.7 25.7 13.0 24.2 25.7
2002-01 -2.0 24.6 17.2 25.7 24.5
2002-02 25.8 22.7 18.3 23.8 23.4
2003-01 27.4 26.3 19.6 25.5 25.6
2003-02 24.0 24.1 16.4 23.6 27.8
```

Unreferenced Data Format

Unreferenced or index data files organise data in the same way as station datasets, i.e. the rows represent time, and the columns the indices. The only differences are that only one field is permitted, and the latitude and longitude lines (and any additional information lines) are omitted. The tags are very similar to those for station data. Compulsory tags are:

* cpt:T=time period for which the data are valid (in the ISO8601 format described above)
* cpt:nrow=number of time steps
* cpt:ncol=number of stations
* cpt:row=T (representing time)
* cpt:col=index

The number of time steps is usually the number of years of data available, but if there are EOF extensions, the number of time steps will increase. For example, if data are available for January and February 1971 to 2000 there are two months (one EOF extension) and thirty years of data, making a total of 60 time steps. Optional tags are:
Input Data Format

* cpt:field=abbreviated generic name for the indices
* cpt:units=units in which the data are stored
* cpt:missing=missing value flag

Immediately beneath the tag_line should be a line containing the names, or some kind of unique identifying symbol, for each of the ncol indices. The name for each index must not be longer than 16 characters, and should contain no spaces. An example is shown below.

xmlns:cpt=http://iri.columbia.edu/CPT/v10/
cpt:nfields=1

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960-01</td>
<td>-2.0</td>
<td>159.4</td>
<td>306.9</td>
<td>35.1</td>
<td>167.4</td>
</tr>
<tr>
<td>1960-02</td>
<td>590.5</td>
<td>53.4</td>
<td>237.0</td>
<td>113.4</td>
<td>411.3</td>
</tr>
<tr>
<td>1961-01</td>
<td>118.5</td>
<td>10.5</td>
<td>37.4</td>
<td>0.0</td>
<td>-2.0</td>
</tr>
<tr>
<td>1961-02</td>
<td>61.0</td>
<td>15.2</td>
<td>69.0</td>
<td>0.0</td>
<td>34.3</td>
</tr>
<tr>
<td>1962-01</td>
<td>-2.0</td>
<td>2.3</td>
<td>194.5</td>
<td>3.3</td>
<td>98.6</td>
</tr>
<tr>
<td>1962-02</td>
<td>130.9</td>
<td>3.6</td>
<td>41.2</td>
<td>0.0</td>
<td>53.3</td>
</tr>
<tr>
<td>1963-01</td>
<td>241.9</td>
<td>2.8</td>
<td>213.7</td>
<td>9.6</td>
<td>141.1</td>
</tr>
<tr>
<td>1963-02</td>
<td>86.1</td>
<td>2.2</td>
<td>108.5</td>
<td>4.3</td>
<td>237.0</td>
</tr>
</tbody>
</table>

Older File Formats (File Format before version 10)

1. Old Gridded Data Format

Old CPT file formats can be read by CPT version 10. Their structure is simpler than the new formats, and so some users may prefer to continue to use the old format. Gridded data files must be structured such that columns represent longitudes, and rows represent latitudes. The date, the latitudes and the longitudes must also be provided. The first row of data must indicate the date of the first record in the format dd mmm yyyy, followed by the longitudes. The second record must contain the latitude followed by the data. Latitudes must be represented as between 90 and -90, with negative values indicating latitudes in the southern hemisphere. It is more slightly efficient to order the latitudes from north to south, but the reverse ordering can be used. Longitudes must be represented as values between -180 and +360, with negative values indicating longitudes in the western hemisphere, and the longitudes must be ordered from west to east. Hence, a portion of a 2.5 degree gridded input file should look something like the following:

01 Aug 1968 0.000000 2.500000 5.000000 7.500000
88.7500 2854.78 2855.14 2855.50 2855.85
86.2500 2854.08 2855.04 2855.96 2856.86
83.7500 2856.63 2858.27 2859.79 2861.19

2. Older Station Data Format

Station data files must be structured such that each column represents a different station, and each row represents the observed values for a given year. However, the first three rows must include the station name or label (maximum of 16 characters with no space in the middle of a name of label), latitude and longitude, respectively, of each station, and must be preceded by the keywords "Station" (or "Stn"), "Latitude" (or "Lat"), and "Longitude" (or "Long" or "Lon"). The keywords are not case-sensitive, and so can be in upper, or lower, or mixed case. The first column of the data file must represent the year, listed consecutively beneath the keywords. (For applications in which the data do not represent years
Input Data Format

it would normally be appropriate to set this column to the record number.) Hence, a valid station input file should look something like the following:

<table>
<thead>
<tr>
<th>STN</th>
<th>STN_A</th>
<th>STN_B</th>
<th>STN_C</th>
<th>STN_D</th>
<th>STN_E</th>
<th>STN_F</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAT</td>
<td>10.0</td>
<td>12.2</td>
<td>11.5</td>
<td>8.2</td>
<td>9.4</td>
<td>10.9</td>
</tr>
<tr>
<td>LONG</td>
<td>-1.0</td>
<td>12.3</td>
<td>-10.2</td>
<td>9.1</td>
<td>-5.3</td>
<td>8.4</td>
</tr>
</tbody>
</table>

1965 0.44 0.09 0.29 0.03 0.06 0.43
1966 0.18 0.34 0.81 1.96 1.09 0.51
1967 1.57 1.00 1.07 1.19 1.61 1.13
1968 0.77 0.99 0.35 0.76 0.52 0.29

3. Older Unreferenced (index) Data Format

Unreferenced data files are similar to the station data files, but do not have the first second and third rows, and use the "Name" keyword (again not case sensitive) instead of the "Station" keyword. ("Year" and "Years" are acceptable alternative keywords.) Thus, unreferenced data files should be structured such that each column represents a different series, and each row represents the observed values for a given year. The first row of the file should indicate the name of each series (maximum of 16 characters with no space in the middle of a name of label). The first column of the data file must represent the year, listed consecutively beneath the "Name" keyword. (For applications in which the data do not represent years it would normally be appropriate to set this column to the record number.) Hence, a valid unreferenced input file should look something like the following:

<table>
<thead>
<tr>
<th>NAME</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1965</td>
<td>0.44</td>
<td>0.09</td>
<td>0.29</td>
<td>0.03</td>
<td>0.06</td>
<td>0.43</td>
</tr>
<tr>
<td>1966</td>
<td>0.18</td>
<td>0.34</td>
<td>0.81</td>
<td>1.96</td>
<td>1.09</td>
<td>0.51</td>
</tr>
<tr>
<td>1967</td>
<td>1.57</td>
<td>1.00</td>
<td>1.07</td>
<td>1.19</td>
<td>1.61</td>
<td>1.13</td>
</tr>
<tr>
<td>1968</td>
<td>0.77</td>
<td>0.99</td>
<td>0.35</td>
<td>0.76</td>
<td>0.52</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Since version 6 of CPT it is no longer permissible for the gridpoints, stations or series to run over to the following line.
Chapter 4. Downloading CPT Data From IRI Data Library

CPT Input Data can be easily created using Microsoft Excel. The IRI/LDEO Climate Data Library (http://iridl.ldeo.columbia.edu/) also offers over 300 freely available climate related datasets than can be easily downloaded in CPTv10 format but also can be downloaded from the IRI Data Library. We highly recommend you to read the IRI Data Library tutorial, which can be accessed by clicking on the 'Tutorial' link under Help Resources in the blue box on the left side of the page Data Library home page. The CPTv10 file format documentation page is also a useful resource http://iridl.ldeo.columbia.edu/dochelp/Documentation/details/index.html?func= cptv10

The next section describes how to download sea surface temperature data in CPT version 10 format.

Downloading Sea Surface Temperature Data

Step 1: Launch web site http://iridl.ldeo.columbia.edu/ []

Step 2: Click 'Finding Datasets' link, the first link of the left blue box. Then click 'By Category' link, the third entry on web page.

Step 3: Click on the first category, 'Air-Sea Inteface'

Step 4: Scroll down and selected the dataset named 'NOAA NCDC ERSST'

Step 5: Select 'version 2'.

Step 6: Select 'Sea Surface Temperature'

You will see the following page.
Figure 4.1. ERSST Data Selection

Step 6: Click 'Data Selection' on the second row of top right table.

You will see a page as following.
Figure 4.2. Data Selection

Step 7: Specify the data range.

There are several ways to specify the data range. One is to use the 'Data Viewer'. Another option is to specify the lat/lon/time range in the second grey box, such as in the example below.
Figure 4.3. Range Selection

Type in Latitude range as 51N to 29N, Longitude range from 125W to 75W and Time series from Jan 1980 to Nov 2009. Click 'Restrict Range'. You will see the information refreshed on the first grey box. You have selected grid data of Y from 50N to 30N and X from 124W to 76W. The data are in 2 degree resolution so the your specified ranges have been automatically adjusted.

You can also use 'Data Viewer' to select the data range. Please read the Data Library tutorial for more information on this.

Step 8: Click 'Stop Selecting' in the first grey box. This will redirect you back to the previous page with data range selection.

Step 9: Click 'Data Files' on the second row of the top right box. Then click 'CPT'. You will see the following.
On this page, the data to be downloaded is described in the table followed by several links allowing you to download data in different CPT formats. Click either link under CPTv10 to download the data file in the latest CPT version format.

**Downloading ECHAM4.5 Forecast Data**

The user should use 'Expert Mode' instead of regular 'Data Selecting' to download analysis data for CPT Version 10.xx.


Step 2: Click 'Finding Datasets' link which is the first link of the left blue box. Then click 'By Category' link which is the third entry on web page.

Step 3: Click 'Forecasts'.

Step 4: Select 'IRI FD ECHAM4.5 Forecast psst ensemble12 MONTHLY' as 'Dataset Name'

Step 5: Select 'Surface'.

Step 6: Select 'Total Precipitation'

Step 7: Select 'Expert Mode' on the first row of top right box. You will see the following.
Figure 4.5. ECHAM Expert Mode

![Figure 4.5. ECHAM Expert Mode](image)

There is a scrollable box on the top left corner. You can specify the data domain in Ingrid language. Here is an example for a precipitation analysis without an explicit time range:

```
expert
SOURCES .IRI .FD .ECHAM4p5 .Forecast .psst .ensemble12 .MONTHLY .surface .prcp
(mm/day) unitconvert
[M]average
S (1 Dec 1957–2008) VALUES
L 0.5 2.5 RANGEEDGES
[L] /keepgrids average
[X Y] [S L add] cptv10
```

The above Ingrid code request to the IRI Data Library dataset selects the precipitation record, changes its units, calculates an ensemble average, selects Dec starts, and averages over the first three lags (retaining the lag (L) grid so that it can be used to compute time later). The last line then specifies X Y as the spatial grids and the sum of start and lead time as time. More examples demonstrating how to query the IRI Data Library using Ingrid code are available at the link [http://iridl.ldeo.columbia.edu/dochelp/Documentation/details/index.html?func= cptv10](http://iridl.ldeo.columbia.edu/dochelp/Documentation/details/index.html?func=cptv10).

Please copy and paste the above Ingrid code to expert mode box and then click OK. You will be redirected to the following page.
Figure 4.6. ECHAM Expert Mode Data File

Step 8: Click 'tsv' or tsv.gz' to download data in CPT Version 10 format. Please refer to documentation from previous versions of CPT for details on older CPT file formats and how to download them.

Download Dual Field Station Data

The user should use 'Expert Mode' instead of regular 'Data Selecting' to download station data for CPT Version 10.xx.

Step 1: Launch web site http://iridl.ldeo.columbia.edu/

Step 2: Click 'Finding Datasets' link which is the first link of the left blue box. Then click 'By Source' link.

Step 3: Click 'NOAA'.

Step 4: Select 'NCDC'

Step 5: Select 'United States Historical Climatology Network'.

Step 6: Select 'Expert Mode' on the first row of top right box. You will see the following
Figure 4.7. Dual Fields Station Data

Step 7: Copy and paste the following Igrid code to the expert query box on the top left corner

```sql
expert
SOURCES .NOAA .NCDC .USHCN
state
(30) dup masknotrange
SELECT
T (Jun 1960-1979) VALUES
{raw .prcp
raw .mean .temp
lon lat elev Name}[ID][T] cptv10
```

This example uses the state code to pick out a subset of the stations, picks June 1960-1979.

Step 8: Select 'OK' on the first row of top right box. You will see the following.
Figure 4.8. Dual Fields Station Data Info

Step 9: Click 'tsv' or tsv.gz’ to download data in CPT Version 10 formats. Again, Please refer to documentation from previous versions of CPT for details on older CPT file formats and how to download them. Both precipitation and temperature fields are included in this data download example.
Part II. CPT Windows Version Tutorial
Table of Contents

5. Getting Started ......................................................................................................................... 24
6. Selecting Analysis Method ....................................................................................................... 25
   Choose the analysis to perform: PCR, CCA or MLR ................................................................. 25
7. Select Input Datasets ............................................................................................................... 26
   Select Input Datasets .............................................................................................................. 26
8. Set Program Parameters .......................................................................................................... 27
   Specify X domain and EOF options ......................................................................................... 27
   Specifying Y Domain and EOF options ................................................................................... 27
   Setting the Training Period, Cross-Validation Window ......................................................... 27
   Setting Missing Values ........................................................................................................... 28
   Set Other Options .................................................................................................................. 28
   Save Program Parameters ...................................................................................................... 29
9. Performing the Analysis ........................................................................................................... 30
   Analysis Progress .................................................................................................................... 30
10. CPT Results ............................................................................................................................ 31
    CPT Graphics ......................................................................................................................... 31
    Saving the Graphics .............................................................................................................. 31
    Customising the graphics ...................................................................................................... 32
    Saving Results to Files .......................................................................................................... 32
11. CPT Results Analysis ............................................................................................................. 34
    ........................................................................................................................................... 34
12. CPT Forecasts .......................................................................................................................... 35
    Selecting Forecast File ........................................................................................................... 35
    Setting Forecast Options ....................................................................................................... 35
    Forecast Results ..................................................................................................................... 36
        Forecast Results: Series .................................................................................................... 37
        Forecast Results: Maps .................................................................................................... 37
        Forecast Results: probabilities of exceedance ................................................................. 37
    Changing Forecast Settings .................................................................................................... 38
Chapter 5. Getting Started

After starting the software, the Introductory Window will be opened, indicating the title of the software, and showing the IRI banner. The Window title (the very top row of the Window), should indicate the version of the software being used. Further details of the software version are available in the menu bar immediately below the title bar, as described below. There should be three menu items:

• File
• View
• Help

From the Introductory Window, the section called “New” menu item offers only the option to Open a pre-existing Project File, or to Exit the software. Help ~ About gives details about the version of the software, and Help ~ Home provides more general help (these pages). These help pages are available from anywhere within the program.

The most important menu item from the Introductory Window is Chapter 18, View. This item offers a choice of analysis methods:

• Canonical Correlation Analysis (CCA)
• Principal Components Regression (PCR)
• Multiple Linear Regression (MLR)

Any of the methods can be used to construct a model for seasonal climate forecasting. The CCA, PCR or MLR Window will be opened depending on which of these items you select. Once you have selected an analysis method you will be prompted for the various settings and input files to construct the forecasting model.
Chapter 6. Selecting Analysis Method

Choose the analysis to perform: PCR, CCA or MLR

In the main menu, click on View, to show a drop-down menu with three options. Choose canonical correlation analysis, principal components regression, or multiple linear regression as your analysis method.

Figure 6.1. Main Menu
Chapter 7. Select Input Datasets

Select Input Datasets

Two datasets are required by CPT. The first dataset contains the "X variables". These variables are sometimes called "predictors", "independent variables" or "explanatory variables". In the context of MOS applications, the X variables will normally be a GCM output field, such as precipitation or geopotential heights, while in a more traditional model the X variables typically will be something like a set of sea-surface temperature data, or an ENSO index. The X variables are used to predict the variables in the second dataset, which should contain the "Y variables". The Y variables are sometimes called "predictands", "dependent variables", or "response variables". Most frequently the Y dataset contains a set of station seasonal rainfall totals or temperature averages.

The CPT program requires the input files to follow strictly one of three structures (gridded, station, and unreferenced or index). Currently, the input files for each of these structures must be in ASCII (or text) format, although other formats are planned for later releases of the software. New input file formats were introduced in CPT version 10.01 or later. The version 10.xx of CPT continues to be able to read files with old file format, but the old formats are unable to support some of the new features of CPT, including multiple fields, so-called "EOF extensions" (which are typically the inclusion of additional predictors at different lags), and improved date handling. Please see Chapter 3, Input Data Format for data format details.

You can use the 'browse' button to choose 'X variables' or 'Y variables'. See the following figure.

Figure 7.1. Input Data

After clicking the 'browse' button, a file chooser dialog box will pop up. The default data directory is C: \Documents and Settings\user\Application Data\CPT\DATA\ but you can select any other directory where your data are located.
Chapter 8. Set Program Parameters

For gridded and station datasets, CPT lets you choose the spatial domain over which you want to perform your EOF or CCA analysis. In general the domain is known in advance through experience. User needs to specify start year, length of training period, length of cross-validation window etc. parameters correctly in order to get meaningful CPT results.

Specify X domain and EOF options

After you select a dataset, you will see a dialog box as following to let your choose your domain of interest. Click 'Data Limits' to choose maximum available domain for the data or specify your domain of interest. Then Click 'OK'. You will see another dialog box to allow you set X EOF settings. See figures below. You have to choose the number of EOFs for the predictor fields used to fit the model. If you set the minimum to be less than the maximum, CPT will find the optimum number of modes between the two numbers. However, if you set the minimum equal to the maximum, then CPT will use that number of modes.

You can modify X or Y domain by clicking 'Edit'=>'X Data Domain' or 'Y Data Domain'. You can also modify EOF options by clicking 'Options'=>'X EOF Options' or 'Y EOF Options'.

Figure 8.1. X Domain Selection

Figure 8.2. X EOF Options

Specifying Y Domain and EOF options

After choosing a Y input file, dialog boxes will appear similar to the X input data dialog boxes. Repeat the above steps.

Setting the Training Period, Cross-Validation Window

By default CPT usually starts the analysis from the first years in the X and Y files; note that these years could be different. You would normally set them equal to the latest year in the two files. You should make sure the lag is correct if you cross the calendar year, for example, while using the DJF or JFM season. In this case the starting year for file X may need to be one year earlier than for file Y. You have to specify the length of the training period as well as the length of the cross-validation window.
Figure 8.3. CCA Settings

Setting Missing Values

If you have missing values in your dataset, you need to specify what you want CPT to do with them.

Click 'Options'--->'Data'--->'Missing Values', to display the Missing Values dialog box as following. In the Missing value flag box, specify the number in your dataset that represents a missing value for the dataset with older file format. The dataset with V10 file format includes the missing value field and CPT will read the value and automatically set it to that value in the file. You can choose the maximum % of missing values. If a station has more than that percentage of missing values, CPT will not use that station in its model. You can also choose which method you want CPT to use to replace the missing values.

Figure 8.4. Missing Values

Set Other Options

On main menu, click 'Options' to display the submenu. This submenu allows the user to change other parameters.
Save Program Parameters

Once you have selected the input files and your settings, you can save these settings in a project file to recall them later: Click File => Save By default, CPT saves all the project files in the subdirectory C:\Documents and Settings\user\Application Data\CPT\Projects\
Chapter 9. Performing the Analysis

In the main menu, click 'Actions' to display the submenu. This submenu allows the user to perform different analyses. See figure below.

Analysis Progress

After the user clicks 'Actions' => 'Calculate' => 'Cross-validated', CPT begins the specified analysis, and the user can view the steps of the analysis and of the optimization procedure right below the setting blocks. The figure below shows setting and progress information after CPT finished a CCA analysis.
Chapter 10. CPT Results

CPT visualizes many of its results. Its graphics make it easier for the user to understand the analysis. The user can save those graphics as image files which can be used for many other purposes. Users can also save results as text files with CPT v10 format.

CPT Graphics

All CPT graphics are generated using the 'Tools' menu. For example, clicking 'Tools'=>'Modes'=>'Scree Plots' to draw 'Scree plots' will generate a graphic such as the one in the figure below, which shows the percentage of variance associated with each EOF plotted.

![Scree Plot](image)

Saving the Graphics

Right click with the mouse on any graphic, and a CPT submenu will appear, allowing the graphic to be saved as an image file in JPEG or BMP format. See the figure below.
Customising the graphics

Options for customizing graphics can also be accessed by right clicking on any graphic. Plot titles can be changed by clicking 'Customise'=>’Title’ and line styles can be changed by clicking ‘Customise'=>’Broken Stick’, etc. See the above figure.

Saving Results to Files

Most CPT results can be saved as text files. Some results can only be displayed as a graphic. Clicking ‘File'=>’Output Files’ shows a tabbed dialog box as below. User can specify file names using ‘browse’ button for all or selected output data of interest, then clicking ‘OK’ to save them.
CPT Results

Output files

- EOF Prefiltering
  - X Eigenvalues: [browse]
  - X Spatial Loadings: [browse]
  - X Temporal Scores: [browse]
Chapter 11. CPT Results Analysis

CPT generates many results. The following sections provide some examples.

From the main menu, clicking on “Tools”:=>’Validation’=>’Cross-Validate’ will show skill, hindcasts and observed series. 'Contingency Tables” shows contingency tables. 'Graphics' show the EOFs time series, loading patterns and scree plots.
Chapter 12. CPT Forecasts

Selecting Forecast File

Once your model is built, you can make a forecast using a forecast file with new records of the X variables. By default CPT selects the same input predictor file. You can change it by clicking the 'browse' button in 'Forecast Variables' block. The forecast file must be in the same format as X file. For example, if the X input file was gridded data, the forecast data file should be gridded data also.

You then select:

Setting Forecast Options

From the 'Main Menu' window 'Forecast Variables' block, user can set

(a) the starting year of the forecasts --> 'Start at:'
(b) the number of years to forecast --> 'Number of forecasts:'

See the figure below.
Forecast Results

After forecast options are set, the user can click 'Tools' => 'Forecasts' to choose different forecast graphic results. See the figure below.
Forecast Results: Series

Click 'Tools'=> Forecast'=> Series', to display the following figure.

This shows the predicted values (green crosses) for the current station as well as forecast possibilities, confidence limits for the forecast and, in the “Thresholds” box, the “category thresholds” as well as the climatological probabilities for the 3 categories.

Forecast Results: Maps

The forecast probabilities map lists the probabilities for each category at each location as well as the spatial distribution of the probabilities.

Figure is coming soon./*FIXME*/

Forecast Results: probabilities of exceedance

To draw the probabilities of exceedance go to: Tools => Forecast => Exceedances
Changing Forecast Settings

User can change forecast settings through 'Options' menu. See below.
Clicking 'Tools'=>'Tailoring' will get the following figure
**Threshold Settings**

### Standardization:
- Select method:
  - No standardization
  - Anomalies
  - Standardized anomalies

### Thresholds:

#### Probabilities:
- Climatological probabilities of outer categories:
  - Above: 0.333333
  - Below: 0.333333

#### Absolute thresholds:
- Absolute thresholds:
  - Upper: 200
  - Lower: 100

### Analogues:
- Analogues:
  - Analogue 1: 1979
  - Analogue 2: 1979
After the settings are changed and the forecast result graphic redrawn, the thresholds or above normal or below normal values, etc. will also change.
Part III. Specific CPT Menu Items
Table of Contents

13. File .................................................................................................................. 46
   New ............................................................................................................... 46
   Open ............................................................................................................ 46
   Save ............................................................................................................ 46
   Save As ....................................................................................................... 46
   Output Results ............................................................................................ 46
   Exit .............................................................................................................. 47
   Close .......................................................................................................... 47
14. Edit .................................................................................................................. 48
   X Data Domain ............................................................................................ 48
   Y Data Domain ............................................................................................ 48
15. Actions ............................................................................................................ 49
   Calculate .................................................................................................... 49
      Cross-Validated .................................................................................... 49
      Retroactive ............................................................................................ 49
   Reset .......................................................................................................... 50
16. Tools ............................................................................................................. 51
   Validation ................................................................................................... 51
      Cross-Validated .................................................................................... 51
      Retroactive ............................................................................................ 52
      Verification ........................................................................................... 52
      Contingency Tables ............................................................................. 54
      Modes ................................................................................................. 54
      Forecasts .............................................................................................. 56
17. Options .......................................................................................................... 60
   X EOF Options ........................................................................................... 60
   Y EOF Options ........................................................................................... 60
   CCA Options .............................................................................................. 60
   Climatological Period ................................................................................ 60
   Tailoring ..................................................................................................... 60
   Data ........................................................................................................... 61
      Transform Y Data ................................................................................ 61
      Zero-Bound ......................................................................................... 61
      Missing Values .................................................................................. 62
   Resampling Settings ............................................................................... 62
   Forecast Settings ..................................................................................... 62
   Graphics .................................................................................................... 62
      Crosses on Graphs .............................................................................. 62
      Mask Land ............................................................................................ 62
      Mask Lakes .......................................................................................... 63
      Black and White .................................................................................. 63
      Reverse Colours .................................................................................. 63
      Vertical Lines on Graph ...................................................................... 63
18. View .............................................................................................................. 64
   Canonical Correlation Analysis (CCA) ....................................................... 64
   Principal Components Regression (PCR) ..................................................... 64
   Multiple Linear Regression (MLR) ............................................................... 64
19. Customize ...................................................................................................... 65
   Scree Plot .................................................................................................. 65
   Cumulative ................................................................................................. 65
   Logarithmic Axis ...................................................................................... 65
Specific CPT Menu Items

Broken Stick .................................................................................................................. 65
Title ............................................................................................................................... 65
Chapter 13. File

New

**Restoring default settings.** Use the New option under the File menu item to restore all default settings. This option will close any opened Project File.

Open

**Accessing Old Project Files.** The program settings, including all input files, can be saved in a Project File. By default these are store in the PROJECTS subdirectory. Open will prompt you for the name of the Project File you wish to access. The File name can either be typed in at the "File name" prompt, or the File can be selected from the list using the mouse.

Note that the Project Files for Canonical Correlation Analysis, Principal Components Regression, and Multiple Linear Regression are all different. The program will by default provide a list of only the appropriate files. It is a mistake to try to read a Project File of the wrong type. However, common settings are maintained when you switch between CCA, PCR and MLR when you use the View menu item. So, for example, if you want to use PCR settings saved in a Project File to perform CCA, open the PCR Project File, and then use View to switch to CCA. Finally, complete any new fields required by CCA.

Save

**Saving settings.** The program settings, including the input file definitions, can be saved in a Project File. Save will store the current settings in the currently opened Project File. If a Project File is not currently open **Save As** is invoked automatically.

The name of the Project File will appear in the title of the CPT Window.

Save As

**Saving settings.** The Canonical Correlation Analysis and Principal Components Regression options can be saved in a Project File. By default these files are stored in the PROJECTS subdirectory of the directory where CPT was installed. Save As will first prompt for a Project File name, and will then store the current settings in the this.

Output Results

Most of the results can be saved in files by using the **Output Results** option. This option becomes accessible only after the calculations are complete. Most results can be saved from here, although the various forecast outputs are only available if a forecast has already been calculated. Verification results, and skill maps are available only by right-clicking on the respective results windows because these are stored in temporary memory.

File names for each of the desired outputs need to be specified. By default these files are stored in the OUTPUT subdirectory. The formats follow standard FORTRAN formatting conventions, except that formatted direct access is not permitted. These conventions are not discussed in these Help Pages. It is recommended that the default settings of formatted sequential output be used to obtain standard text (ASCII) files in which fields are separated by tabs. Formatted output files are given the .txt file extension, and are saved in CPT version 10 format, while unformatted files are given the .dat extension.
Exit

Exiting CPT To exit CPT use the Exit option under the File menu. If a Project File is open and the current settings have not been saved, the user will be prompted to save the Project.

Close

To close a window without exiting CPT, use the Close option under the File menu.
Chapter 14. Edit

X Data Domain

It is possible to select a rectangular sub-area of a gridded/station input dataset. The domain is specified by setting its latitudinal and longitudinal limits. Once set, the number of gridpoints in a gridded data analysis will be automatically reset on the CPT Window. For a station dataset, the number of stations on the CPT Window will continue to show the total number of stations in the dataset as originally specified. This menu item is available only if the input dataset is not defined as unreferenced. The program automatically prompts for the domain settings when a geographically referenced (gridded or station) input file is opened, but this menu item allows the domain to be reset without having to re-open the file. Note that changing the X domain will require CPT to reset if results have been calculated.

Y Data Domain

It is possible to select a rectangular sub-area of a gridded/station input dataset. The domain is specified by setting its latitudinal and longitudinal limits. Once set, the number of gridpoints in a gridded data analysis will be automatically reset on the CPT Window. For a station dataset, the number of stations on the CPT Window will continue to show the total number of stations in the dataset as originally specified. This menu item is available only if the input dataset is not defined as unreferenced. The program automatically prompts for the domain settings when a geographically referenced (gridded or station) input file is opened, but this menu item allows the domain to be reset without having to re-open the file. Note that changing the X domain will require CPT to reset if results have been calculated.
Chapter 15. Actions

Calculate

The Calculate menu item becomes available only when both X and Y input datasets have been specified. Once specified, the CCA, PCR, or MLR model can then be fit, and validation statistics calculated, using one of two options: Cross-Validated or Retroactive.

Cross-Validated

The cross-validated calculation option fits a CCA, PCR, or MLR model using all data within the training period. This model is used to make any forecasts using predictor data supplied in the "forecast file". The option also produces cross-validated forecasts for each year in the training period. (Note that these cross-validated forecasts, and therefore the available validation statistics, are purely deterministic: no uncertainty estimates are generated. If a set of past probabilistic forecasts is required, perhaps with accompanying verification statistics, the retroactive option should be selected.) At each cross-validation step k consecutive years are omitted from the training period, where k is the "length of the cross-validation window". The model is then completely reconstructed, including recalculating the principal components, and redefining the category thresholds, and the middle year of the years omitted from the training sample is forecast. This process is repeated for each year. Towards the beginning and end of the training period the cross-validation window is looped to ensure that exactly k years are always omitted. For example, if k = 5, when forecasting the first year, the first three years are omitted together with the last two; when forecasting the second year, the first four years are omitted together with the last one.

The cross-validated forecasts are made available for output to a file, and for performance analyses within CPT. Note that all the information regarding the definitions of the principal components and CCA modes (where applicable) is based on the results using all the data within the training period, but all the results regarding the performance of the model (validation) are based on the cross-validated forecasts.

Retroactive

The retroactive calculation option fits a CCA, PCR, or MLR model to an initial subset of the training period that must be specified. The initial subset consists of the first few years of the training period. The default is to use the first half of the full training period, but the user may wish to define a more appropriate initial training period. The shorter the initial training period is, the more years are available for producing retroactive forecasts, and the more reliable the performance measures are likely to be. However, if the initial training period is made too short, sampling errors in constructing the models may result in inaccurate retroactive forecasts.

After fitting a model to the initial training period, cross-validated forecasts are made for each year in this initial shortened training period, and retroactive forecasts are then produced for the next k years, where k is the model update interval. The model is then refitted using the initial training period plus these first k years, a new set of cross-validated predictions is made, and retroactive forecasts are made for the following k years. This procedure is repeated until the last year of the full training period is forecast. These forecasts are produced deterministically and probabilistically. The probabilistic forecasts are calculated based on the error variance of the most recently available cross-validated predictions.

The retroactive procedure then fits a model using the full training period, and cross-validates this model. These cross-validated forecasts are identical to those that are obtained using the cross-validated option. The cross-validated and the retroactive forecasts are available for performance analyses, and can be saved to an output file. (Note that only the cross-validated predictions for the entire training period are available: the preliminary cross-validated predictions for the shortened training periods are not retained.)
If the model is to be optimized, by identifying the best numbers of EOF and/or CCA modes, an optimal model is identified each time the retroactive model is updated. As a result, the retroactive option with optimization may take a while to compute all the forecasts, especially if the model update interval is short.

**Reset**

The reset option erases all results from memory, and closes any open results windows, but leaves all input files and program settings unchanged. It is necessary to reset CPT if certain program settings are to be modified, and that would affect the fitted model. For example, redefining the length of the training period changes the sample data used to construct the model, and so all results need to be recalculated. However, changing the number of forecasts does not require a reset since the same model is used regardless of which and how many forecasts are to be made. CPT will automatically prompt the user if modifying a program setting would require a reset, but some menu items are deactivated (for example, setting a zero bound) and so the user will need to reset CPT manually using the reset menu item.

The reset option becomes available only after cross-validated or retroactive forecasts have been generated.
Chapter 16. Tools

Validation

Validation scores and graphics can be obtained for the cross-validated forecasts, and if the retroactive forecast option was selected then verification for these forecasts is also available. Select the forecasts that are to be verified, and a validation window will open.

Cross-Validated

A set of scores and graphics can be obtained for the cross-validated forecasts for a single series, or a map (if the Y data are in gridded or station format) or bar chart (otherwise) for all series can be displayed.

Performance Measures

The validation window provides a variety of forecast performance scores together with a graph of forecast (green line) and observed (red line) values, and ROC graphs for the above- (red line) and below-normal (blue line) categories. The categorical measures and the ROC graphs are based on three categories (see Contingency Tables for further details on the definitions of the categories). Scores and graphs are shown for one series at a time. Verification for the desired series can be shown by setting the appropriate number at the top left of the validation window. A series that has been omitted in the calculations will be skipped.

The graphics can be saved as JPEG files by right clicking anywhere in the child window, and then selecting the desired graphic to be saved. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. The graphics titles can be set using the Customise option upon right-clicking in the child window.

Further details are available under Viewing results.

Bootstrap

The bootstrap window provides confidence limits and significance tests for a variety of forecast performance scores. The confidence limits are calculated using bootstrap resampling, and provide an indication of the sampling errors in each performance measure. The bootstrap confidence level used is indicated, and can be adjusted using the Options ~ Resampling Settings menu item. The actual sample scores are indicated, and are the same as those provided by Performance Measures.

As well as providing confidence limits, significance levels are also provided. The p-value indicates the probability that the sample score would be bettered by chance. Permutation procedures are used to calculate the p-values. The accuracy of the p-values depends upon the number of permutations, which can be set using the Options ~ Resampling Settings menu item. It is recommended that at least 200 permutations be used, and more if computation time permits.

Skill Map/Bar Chart

The validation window provides a map (if the Y data are in gridded or station format) or a bar chart (otherwise) of a specific performance measure for all series at once. Graphics for the desired score can be shown by selecting the appropriate button.
Tools

The graphic can be saved as a JPEG file by right clicking anywhere in the child window. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. The graphic title can be set using the Customise option upon right-clicking in the child window.

Further details are available under Viewing results.

Scatter Plots

The Scatter Plots option shows a graph of the forecast residuals (differences between the forecasts and the observations), as well as a scatter plot of the observations against the forecasts. The scatter plot includes horizontal and vertical divisions that indicate the three categories. In both cases the divisions are defined by the terciles of the observations using all the cases. A least-squares linear regression line is shown on the scatter plot, but only over the range of the forecasts.

The graphics can be saved as JPEG files by right clicking anywhere in the child window. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. The residuals and scatter plots titles can be set using the Customise option upon right-clicking in the child window.

Further details are available under Viewing results.

Retroactive

A set of scores and graphics can be obtained for the retroactive forecasts for a single series, or a map (if the Y data are in gridded or station format) or bar chart (otherwise) for all series can be displayed.

Duplicate validation options above or reorganize?

text here

Verification

Verification scores and graphics can be obtained for the retroactive forecasts. Specifically attributes (or reliability) diagrams, ROC graphs, and some "Weather Roulette" graphics (cumulative profits and effective interest rate graphs) can be drawn. These verification results are based on the retroactive probabilistic forecasts (evaluation of the quality of the deterministic retroactive forecasts is available from the Validation) menu item. The verification results become unavailable if the absolute category thresholds are activated and the thresholds or the standardization option are not adjusted.

Attributes Diagram

The attributes diagrams are constructed using probabilistic retroactive forecasts. The reliability statistics are calculated using predictions for all gridpoints / stations / series. The probabilistic forecasts are binned to the nearest 10% (in accordance with the WMO’s Standardized Verification System for Long-Range Forecasts - SVSLRF), but the point on the x-axis is shown at the average of the forecasts in the bin rather than at the centre of the bin. Results are shown separately for the three categories, and for all the categories pooled together (top left).
The reliability curve is shown by the thick line, while the solid 45 degree diagonal shows the line of perfect reliability. The dashed diagonal is the so-called no-skill line (a reliability curve that is shallower than this will score a negative Brier skill score against climatology). The horizontal line represents the line of no resolution and crosses the y-axis at the observed relative frequency for all years. Note that his value is not necessarily the same as the climatological probability of an event since the observed relative frequency is defined only over the retroactive period.

The reliability statistics can be saved by right-clicking on the graphs. In a similar manner the graphs themselves can be saved and the titles customised.

**ROC Graphs**

The ROC graphs are constructed using probabilistic retroactive forecasts. The ROC statistics are calculated using predictions for all gridpoints / stations / series. The probabilistic forecasts are binned to the nearest 10% (in accordance with the WMO's Standardized Verification System for Long-Range Forecasts - SVSLRF). Results are shown separately for the three categories, and the areas beneath the graphs, calculated using the trapezium rule, are shown.

The ROC statistics can be saved by right-clicking on the graphs. In a similar manner the graphs themselves can be saved and the titles customised.

**Weather Roulette**

Weather Roulette is a suite of verification procedures aimed at providing an indication of the potential value of forecasts. The underlying assumption is that a user of the forecasts invests an initial sum of money on the forecast, apportioning the investment across the three possible outcomes according to the forecast probabilities. The investment pays out fair odds against climatology. For example, if the three categories are climatologically equiprobably the investment will return 3 times (1 / 33%) the investment on the verifying outcome. Whatever returns are made after the first forecast are available for the second forecast. If there are many locations, results are averaged across them all. The change in the investors fortunes can be viewed in one of two ways, as a cumulative profits or effective interest rate diagram.

**Cumulative Profits Graph**

The cumulative profits graph shows how an initial investment of $1 apportioned according to the forecast probabilities, would have increased or decreased over the period of retroactive forecasts. The profits are averaged across all locations.

The graph can be saved as a JPEG file by right-clicking anywhere in the child window. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. The graph title can be set using the Customise option upon right-clicking on the graphic. Similarly, the results can be saved by right-clicking.

**Effective Interest Rate**

The effective interest rates graph translates the cumulative profits into an effective interest rate.

The graph can be saved as a JPEG file by right-clicking anywhere in the child window. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. The graph title can be set using the Customise option upon right-clicking on the graphic. Similarly, the results can be saved by right-clicking.
Contingency Tables

Contingency tables can be obtained for the cross-validated forecasts, and if the retroactive forecast option was selected then tables for these forecasts are also available. Select the forecasts that are to be verified, and a contingency table window will open.

Cross-Validated Contingency Tables

The cross-validated contingency table window provides frequency and contingency tables for the cross-validated forecasts. The frequency table gives the counts of forecasts for each of three categories, marked below-normal (B), normal (N), and above-normal (A), and the number of times each of the three categories verified. The contingency table indicates the percentages of times that each of the three categories verified given the forecast category, and can be obtained from the frequency tables by simply dividing each element in the table by the respective column total. The column and row totals indicate the percentages of times that each of the three categories were forecast and observed, respectively.

Further details are available under Viewing results.

Retroactive Contingency Tables

The retroactive contingency table window provides frequency and contingency tables for the retroactive. The frequency table gives the counts of forecasts for each of three categories, marked below-normal (B), normal (N), and above-normal (A), and the number of times each of the three categories verified. The contingency table indicates the percentages of times that each of the three categories verified given the forecast category, and can be obtained from the frequency tables by simply dividing each element in the table by the respective column total. The column and row totals indicate the percentages of times that each of the three categories were forecast and observed, respectively.

Further details are available under Viewing results.

Modes

Scree Plots

The scree plots show the percentage of variance explained by each EOF mode. The percentages are shown for all modes regardless of the numbers that were specified for use. The plots can be useful in selecting the appropriate number(s) of EOF modes to retain. One method for selecting the number of modes involves identifying “elbows” in the scree plot. The number of modes up to, but not including, the elbow are retained on the basis that the additional modes explain similar amounts of variance, which are assumed to be largely noise. Elbows are often easier to identify when the percentage variance is plotted on a logarithmic axis. An option to use logarithmic axes is available by right-clicking on the scree plots. Another simple method for selecting the number of modes is to retain a sufficient number to ensure that the total variance retained exceeds a predefined minimum. The cumulative percentage of variance explained by the EOF modes can be shown using the Cumulative option by right-clicking on the scree plots. Note that the cumulative percentages are shown only if the logarithmic axis option is switched off.

Another method for selecting the number of models is to compare the explained variance of each mode to that expected from random data. A computationally and conceptually simple such approach is based on the random stick theorem, which involves randomly dividing the total variance into segments and considering the expected size of the nth largest segment. The sizes of the random stick segments can be included on the scree plot by an option available by right-clicking on the plot. The random stick is shown only if the cumulative option is switched off.
The graphics can be saved as JPEG files by right clicking anywhere in the child window, and then selecting the desired graphic to be saved. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. The scree plot titles can be set using the Customise option upon right-clicking in the child window.

**X EOF Loadings and Scores**

If gridded or station data are used, a map of the spatial loadings for each mode is shown, together with the temporal scores. If the data are unreferenced, a bar chart of the loadings is shown instead. The loadings are shown in the form of correlations between the original gridded/station/index data and the temporal scores. If the loadings are saved to an output file, the actual loadings are saved rather than the correlations shown in the map. The graphics for each mode can be cycled by using the arrows at the top left of the graphics window, or by typing in the requisite number. Depending on your computer speed, you may need to wait a short time for the new graphics to display. The graphics can be saved as JPEG files by right clicking anywhere in the child window, and then selecting the desired graphic to be saved. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. The map and graph titles can be set using the Customise option upon right-clicking in the child window.

**Y EOF Loadings and Scores**

If gridded or station data are used, a map of the spatial loadings for each mode is shown, together with the temporal scores. If the data are unreferenced, a bar chart of the loadings is shown instead. The loadings are shown in the form of correlations between the original gridded/station/index data and the temporal scores. If the loadings are saved to an output file, the actual loadings are saved rather than the correlations shown in the map. The graphics for each mode can be cycled by using the arrows at the top left of the graphics window, or by typing in the requisite number. Depending on your computer speed, you may need to wait a short time for the new graphics to display. The graphics can be saved as JPEG files by right clicking anywhere in the child window, and then selecting the desired graphic to be saved. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. The map and graph titles can be set using the Customise option upon right-clicking in the child window.

**CCA Maps**

CCA Maps The CCA maps (or bar charts if the data are unreferenced) show similar results to the EOF Loadings and Scores, but show the weights for the X and Y components of the CCA modes together. The corresponding temporal scores on the centre graph (the red line is for the X variables, and the green line is for the Y variables). It is the correlation between the curves shown on this central graph that CCA attempts to maximize. The canonical correlation is a function partly of any relationships between the X and Y data that CPT exploits to make predictions, but also partly a function of the numbers of X EOF modes retained. Thus the canonical correlation can be artificially inflated simply by forcing COT to retain a large number of EOFs. High canonical correlations should not necessarily be interpreted as a sign of a good predictive model.

The weights and scores for different CCA modes can be displayed using the control at the top left of the CCA maps window, but, unlike the EOF modes, only those CCA modes that CPT retains can be viewed. The results can be saved to an output file, but the actual weights are saved rather than the correlations.
between the original gridded / station / index data and the temporal scores for the corresponding X and Y components of the CCA modes. The weights are known as "homogeneous maps" (even for unreferenced data).

The graphics can be saved as JPEG files by right clicking anywhere in the child window, and then selecting the desired graphic to be saved. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. The map and graph titles can be set using the Customise option upon right-clicking in the child window.

**Forecasts**

Forecast results can be viewed as maps of forecast probabilities or forecast values, or more detailed information is available for the forecasts at individual gridpoints/stations/series.

**Forecast Series**

Tables of the forecasts for individual gridpoints/stations/series and a graph of the historical and current forecast(s) are shown. The table shows the years for which the forecasts apply, some information about the definitions of the categories, and the forecasts themselves, presented in different formats.

In the thresholds box (top left) the climatological period is first indicated, and defines the data used to define the categories. The thresholds dividing the categories are then given in the units of the original data, possibly standardized depending on the standardization option. For example, if the original data are in mm, and the anomaly standardization is selected, the thresholds will be in mm anomalies from the climatological average. The climatological probabilities, below, show the proportion of years in the climatological period that were in each category. By default these probabilities will be 33% (assuming there are no ties, in which case the probabilities may deviate from 33% somewhat), but the actual probabilities will depend upon the threshold settings. The climatological odds are equal to the probability that the category in question was observed, divided by the probability that it was not observed. For example, if the climatological probability of "above-normal" is 33% the odds will be 0.5 ("above-normal" occurs half as often as "normal" and "below-normal" combined, i.e., it occurs half as of often as it does not occur).

The forecasts box presents the forecast in different formats. Probabilistic forecasts and odds are presented first. If the climatological probabilities are equiprobable and the forecast probability for the "normal" category is the lower than for the outer categories, it may be worth transforming the Y data so that they are normally distributed. Applying the transformation will require CPT to reset. The probabilities are derived from the best-guess forecast (listed as "Forecast" under "Forecast ranges") by assuming that the errors in the best-guess forecast will be normally distributed (or, more strictly, follow a Student's t distribution), with the variance of the errors defined by the variance of the errors in the cross-validated predictions (alternative error variances can be set using Options ~ Forecast Settings). If the Y data have been transformed it is assumed that the forecast errors of the transformed data will be normally distributed. From the best-guess forecast and the assumed error variance and distribution the probabilities of exceeding the various thresholds can be calculated. Prediction intervals can also be calculated for a given level of confidence. The prediction intervals are shown as the upper and lower limits under "Forecast ranges". The default is to define the intervals as one standard error from the best guess, which translates to a confidence level of about 68%. The default can be adjusted using Options ~ Forecast Settings.

The graph shows the history of the cross-validated forecasts (green line) and observations (red line), and the current forecast(s) shown by green crosses. The prediction intervals for the forecasts can be added by right-clicking on the graph, and following the "Customize" option. The vertical coloured divisions show the three categories. Note that when calculating the cross-validated forecasts, the threshold definitions will have fluctuated slightly, and so the current categorization fo the forecasts and observations may not exactly
Tools

match the cross-validated classification. This fluctuation is the reason for any apparent discrepancies between the graph and the contingency tables. The graph can be saved as a JPEG file by right-clicking anywhere on the graphic. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. In most cases, the highest quality image should be manageable. The map titles can be set using the Customise option upon right-clicking on the graphic. The forecast values can be saved by using File ~ Output Results.

**Forecast Ensembles**

Tables of an ensembles of forecasts for individual gridpoints/stations/series and a graph of the historical and current forecast(s) are shown. The table shows an ensemble of forecast values for the current variable and forecast target period.

The ensemble is generated by defining equi-probable intervals between the forecasts, based on the cross-validated (by default) error distribution. The default number of ensemble members is nine, so that there is a 10% probability of the observed value falling between and two sorted member, and a 10% probability of the observed value being larger than the largest member, and another 10% probability of it being smaller than the smallest member. The number of ensemble members can be reset using Options ~ Forecast Settings), as can the method for calculating the error variance. The retroactive and fitted error variances are alternative options.

The graph shows the history of the cross-validated forecasts (green line) and observations (red line), and the current ensemble of forecast(s) shown by green crosses. The prediction intervals for the forecasts can be added by right-clicking on the graph, and following the "Customize" option. The vertical coloured divisions show the three categories. The graph can be saved as a JPEG file by right-clicking anywhere on the graphic. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. In most cases, the highest quality image should be manageable. The map titles can be set using the Customise option upon right-clicking on the graphic. The forecast values can be saved by using File ~ Output Results.

**Exceedance Probablities**

Tables of an ensembles of forecasts for individual gridpoints/stations/series and a graph of the historical and current forecast(s) are shown. The table shows an ensemble of forecast values for the current variable and forecast target period.

The ensemble is generated by defining equi-probable intervals between the forecasts, based on the cross-validated (by default) error distribution. The default number of ensemble members is nine, so that there is a 10% probability of the observed value falling between and two sorted member, and a 10% probability of the observed value being larger than the largest member, and another 10% probability of it being smaller than the smallest member. The number of ensemble members can be reset using Options ~ Forecast Settings), as can the method for calculating the error variance. The retroactive and fitted error variances are alternative options.

The graph shows the history of the cross-validated forecasts (green line) and observations (red line), and the current ensemble of forecast(s) shown by green crosses. The prediction intervals for the forecasts can be added by right-clicking on the graph, and following the "Customize" option. The vertical coloured
divisions show the three categories. Note that when calculating the cross-validated forecasts, the threshold definitions will have fluctuated slightly, and so the current categorization fo the forecasts and observations may not exactly match the cross-validated classification. This fluctuation is the reason for any apparent discrepancies between the graph and the contingency tables. The graph can be saved as a JPEG file by right-clicking anywhere on the graphic. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. In most cases, the highest quality image should be manageable. The map titles can be set using the Customise option upon right-clicking on the graphic. The forecast values can be saved by using File ~ Output Results.

**Forecast Maps / Bar Charts**

A set of maps can be drawn to illustrate the forecasts for all grids / stations / series for a given forecast time. The forecasts can be displayed either as actual forecast values (or anomalies, or % departures, etc, based on the standardization setting) with prediction intervals, or as probabilities.

**Forecast Probabilities**

A set of three maps (if the data are gridded or station) or bar charts (otherwise) are shown. Each map indicates the forecast probability for the three categories. The maps are shaded from blue for low probabilities, to red for high probabilities. If more than one forecast is produced, the forecasts can be cycled using the up and down arrows, or typing in the year. The probabilities are derived from the best-guess forecast (the regression estimate from the CCA, PCR, or MLR model) by assuming that the errors in the best-guess forecast will be normally distributed (or, more strictly, will follow a Student's t distribution), with the variance of the errors defined by the variance of the errors in the cross-validated predictions (alternative error variances can be set using Options ~ Forecast Settings). If the Y data have been transformed it is assumed that the forecast errors of the transformed data will be normally distributed. From the best-guess forecast and the assumed error variance and distribution the probabilities of exceeding the various thresholds can be calculated.

The maps can be saved individually as JPEG files by right clicking anywhere in the child window, and then selecting the map to be saved. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. The map titles can be set using the Customise option upon right-clicking in the child window. The forecast values can be saved by using File ~ Output Results.

**Forecast Values**

A set of three maps (if the data are gridded or station) or bar charts (otherwise) are shown. The one map indicates the actual values of the forecasts, while the other two indicate the upper and lower limits of the prediction interval. The values can be shown as anomalies or standardized anomalies by selecting the appropriate option from the Tailoring menu item. The values shown for the predicion intervals depend upon the level of confidence set. See Forecast Settings for details on how to set the confidence level. The prediction intervals are derived from the best-guess forecast (the regression estimate from the CCA, PCR, or MLR model) by assuming that the errors in the best-guess forecast will be normally distributed (or, more strictly, will follow a Student's t distribution), with the variance of the errors defined by the variance of the errors in the cross-validated predictions (alternative error variances can be set using Options ~ Forecast Settings). If the Y data have been transformed it is assumed that the forecast errors of the transformed data will be normally distributed.

If more than one forecast is produced, the forecasts can be cycled using the up and down arrows or typing in the year. The maps can be saved individually as JPEG files by right-clicking anywhere on the graphic,
and then selecting the map to be saved. A default name is given to the file, but this can be changed using the browse button. The quality of the JPEG file can be adjusted using the slider or by the quality indicator, which ranges between 0.01 and 1.00. The highest quality is obtained using 1.00. The size of the JPEG file is affected by the quality chosen, with larger files being generated the higher the selected quality. The map titles can be set using the Customise option upon right-clicking on the graphic. The forecast values can be saved by using File ~ Output Results.
Chapter 17. Options

X EOF Options

Prompts for the maximum and the minimum numbers of X EOF modes are given automatically after opening an X input file when using CCA or PCR, but these values can be reset from the X EOF options menu item. Note that changing the number of EOF modes will require CPT to reset if results have been calculated. For more details on specifying the numbers of modes see EOF modes.

By default the EOFs are calculated using the correlation matrix (the analysis is based on the standardized anomalies). However, it is possible to base the EOFs on either the variance-covariance matrix (the anomalies) or the sums of squares and cross-products matrix (the raw values). To set the method for calculating the EOFs, click on the "Advanced" button, and then choose the required option.

Y EOF Options

Prompts for the maximum and the minimum numbers of Y EOF modes are given automatically after opening an Y input file when using CCA or PCR, but these values can be reset from the Y EOF options menu item. Note that changing the number of EOF modes will require CPT to reset if results have been calculated. For more details on specifying the numbers of modes see EOF modes.

By default the EOFs are calculated using the correlation matrix (the analysis is based on the standardized anomalies). However, it is possible to base the EOFs on either the variance-covariance matrix (the anomalies) or the sums of squares and cross-products matrix (the raw values). To set the method for calculating the EOFs, click on the "Advanced" button, and then choose the required option.

CCA Options

The CCA Options menu permits the user to change the maximum and minimum numbers of CCA modes. A prompt for the numbers of CCA modes does appear automatically after the EOF options are specified if both the X and Y input files are opened, and if the number of CCA modes is not constrained to be one (see comments below regarding restrictions on the numbers of modes). The menu item allows the numbers of modes to be reset at any time. Note that changing the number of CCA modes will require CPT to reset if results have been calculated.

There are a few restrictions on the number of CCA modes: * the minimum number of CCA modes cannot exceed the minimum number of X or Y EOF modes; * the maximum number of CCA modes cannot exceed the maximum number of X or Y EOF modes.

Climatological Period

The climatological period option allows you to specify the period used to calculate the terciles (or other thresholds) used for the forecasts. Forecast categories and probabilities will be calculated using the new climatological period. Cross-validated and retroactive results are unaffected, although the new terciles will be shown on all graphs.

Tailoring

The tailoring options provide a number of ways of providing flexibility in how the forecasts are expressed. The standardization options provide different ways of comparing the forecast to a set of references, and of
defining how the category thresholds are communicated. By selecting the "no standardization" option, the forecasts are expressed in the original units of the input Y data, but these can be converted to anomalies, standardized anomalies, or, if the Zero Bound option is switched on, to percentage departures from average. If the % of average standardization option is selected, and the zero bound option is subsequently switched off, the standardization is switched off (a warning is provided). The forecasts, and saved data, and the category thresholds will be expressed in whichever standardization option is selected.

The tailoring menu option also allows you to specify the prior probabilities of the above- and below-normal categories, or to use thresholds defined in physical values (mm, for example) instead, or to use previous years' observations. By default the prior probabilities are set at 33% so that the boundaries between the categories are the terciles of the climatological distribution, but these probabilities can be reset, thus allowing probabilities for more extreme rainfall events to be calculated, for example. However, it should be noted that errors in estimating the thresholds for extreme events can become large, and so changes in the calculated probabilities may be inaccurate. If the absolute thresholds are set, and the lower threshold is set to be larger than the upper threshold, the two are swapped without warning. The values of the thresholds are assumed to be in the same units as the values of the (possibly standardized) Y input data. If the standardization option is set to "anomalies" and absolute thresholds are used, "below-normal rainfall could be set as 100 mm less than average by setting the lower threshold to -100. The analogue year option sets the larger of the observed values for the two years as the threshold for the "above-normal" category, and the smaller as the threshold for the "below-normal" category. Note that it is possible that the one year may represent the upper threshold for one location but the lower threshold for another location. This option is useful for answering questions such as "What is the probability that there will be more rainfall than last year?" Or "What is the probability that it will be colder than in 1980?"

Data

Transform Y Data

Ideally the Y data should be normally distributed when performing PCR or CCA. The Transform Y Data option attempts to transform the data to a normal distribution before performing the calculations. However, all forecasts are transformed back to the empirical distribution of the data, and all skill calculations (except the goodness index) are performed using the empirical distribution so that the transformation remains largely blind to the user. It should be noted that the output of the Y EOF prefiltering, the various regression coefficients from PCR, and the homogeneous maps of the CCA will apply to the transformed data. Missing data values are estimated using the untransformed data. Selecting the Tranform Y Data option toggles the transformation.

The data transformation proceeds by converting the original data to ranks and expressing these ranks as percentiles of the empirical distribution. Thus, the data are first transformed to a uniform distribution on the unit interval. These percentiles are then used to calculate standard normal distribution deviates. Calculations proceed using the standard normal deviates, and are transformed back to the empirical distribution by first converting to percentiles, and then linearly interpolating the percentiles on the original data. If the percentile lies beyond the range of the original data, then a linear interpolation is applied based on the difference between the most extreme and the second most extreme values.

Note that the transformation may not be effective if there are a number of ties in the original data.

Zero-Bound

For some predictands, such as precipitation, it does not make sense to predict negative values. However, the fact that zero is an absolute lower limit is not considered in the statistical procedures CPT uses. The zero-bound option automatically resets all non-negative predictions (including those for the confidence limits) to zero if they are negative. If there are any negative values in the input data, these are identified.
and an error message is given. If the anomalies or standardized anomalies option is switched on the zero-bound option will still apply, but only at the point where absolute zero is defined in terms of anomalies. (For example, if the climatological mean is 50, and the anomalies option is activated, anomalies of less than -50 will be reset to -50).

**Missing Values**

The Missing Values option allows you to set options for identifying and replacing missing values. A flag should be set to identify missing values in the input datasets. In CPT version 10 files, the missing value flag is typically defined, and so this value should not normally be modified. A maximum percentage of missing values in each series should also be defined. If the number of missing values is less than the defined maximum, these values will be replaced using the method defined.

Replacing Missing Values gives further details about missing value estimation.

**Resampling Settings**

The resampling settings option allows you to define the confidence level and the numbers of resamples used in estimating the confidence intervals and significance levels of the various performance measures. The confidence intervals are calculated using bootstrap resampling, whereas the significance levels are calculated using permutation procedures. The significance levels (or p-values) can be calculated for the cross-validated and retroactive skill maps, and can be saved (but currently not viewed) using the right mouse button on the current map.

**Forecast Settings**

The forecast settings option also allows you to specify the confidence level used when calculating the prediction intervals for new forecasts. If the confidence level is set at p%, then p% of the prediction intervals should contain the observation. The higher the confidence level, the wider the interval will be. Standard settings are 90%, 95%, and 99%. Note that the new confidence level will only take effect when a new forecast window is opened. The method by which the error variance is calculated can be selected.

**Graphics**

**Crosses on Graphs**

The crosses on graph option option allows you to specify whether crosses should appear at the data points on the graphs. The graphs are joined by lines, and will be marked by crosses in addition, if this option is switched on. The option toggles the drawing of the crosses on and off. A tick is indicated next to the menu item when the crosses on graph option is on.

**Mask Land**

The mask land option option allows you to specify whether shading should appear over any areas representing land. For gridded data, CPT creates maps by shading individual grid boxes. When the gridded data represent values such as sea-surface temperatures, the shading of near-coastal gridboxes will extend over land by default. This shading can be prevented by switching on the mask land option; after producing the basic map CPT will then shade all land areas in white so that colours appear only over the sea. The option toggles the masking of the land on and off. A tick is indicated next to the menu item when the masking option is on.
Mask Lakes

The mask lakes option allows you to specify whether shading should appear over any areas representing lakes or inland seas. CPT draws lakes and inland seas on top of the country boundaries, and so these boundaries appear within the lakes and seas themselves. If mask lakes option is on CPT will shade all lakes and inland seas in white so that the borders, and any colour shading will be eliminated. The option toggles the masking of the land on and off. A tick is indicated next to the menu item when the masking option is on.

Black and White

The black and white option allows you to specify whether graphics output should be in black and white. Note that gray-shading is used. The option toggles the black and white graphics output on and off. A tick is indicated next to the menu item when the black and white option is on.

Reverse Colours

The reverse colours option allows you to invert the red-blue colour system, which is appropriate for rainfall, to a blue-red system, which is appropriate for temperature. Note that the colour reversal does not apply to the maps. The option toggles the reversal of the colours on and off. A tick is indicated next to the menu item when the colours are reversed (i.e. oriented for temperature rather than rainfall).

Vertical Lines on Graph

The vertical lines on graph option allows you to specify whether vertical lines should appear on the graphs at regular intervals. If this option is switched on, vertical lines will appear at every five years on the time series graphs, and at every five points on the scree plots. These lines can be helpful in identifying specific data points. The option toggles the drawing of the vertical lines on and off. A tick is indicated next to the menu item when the vertical lines option is on.
Chapter 18. View

There are three analysis methods available in CPT, Canonical Correlation Analysis (CCA), Principal Components Regression (PCR), and Multiple Linear Regression (MLR).

To apply a specific method select the appropriate menu item from View. All three analysis menu items are available only from the Introductory Window.

From the CCA screen only the PCR and MLR menu items are available, and selecting one of these will switch the user to the respective screen. Note that an opened CCA Project File will not be closed, but will become inactive. On returning to the CCA screen, the Project File will be reactivated. However, note that current settings are transferred between analyses. So any changes made to settings on the PCR or MLR screen, will be carried over on returning to the CCA screen. If these new settings are not required, the CCA Project File should be re-opened.

Canonical Correlation Analysis (CCA)

From the Introductory Window the Canonical Correlation Analysis (CCA) menu item will open the CPT Window for CCA. If this menu item is selected from CPT Window when performing PCR or MLR, any open Project Files will be closed, but current settings will be retained.

Principal Components Regression (PCR)

From the Introductory Window the Principal Components Regression (PCR) menu item will open the CPT Window for PCR. If this menu item is selected from CPT Window when performing CCA or MLR, any open Project Files will be closed, but current settings will be retained.

Multiple Linear Regression (MLR)

From the Introductory Window the Multiple Linear Regression (MLR) menu item will open the CPT Window for MLR. If this menu item is selected from CPT Window when performing CCA or PCR, any open Project Files will be closed, but current settings will be retained.
Chapter 19. Customize

Scree Plot

Cumulative

The cumulative percentage of variance can be shown by toggling the Cumulative menu item.

Logarithmic Axis

The y-axis on the scree plot defaults to linear, but it is often easier to identify "elbows" when a logarithmic axis is used. See Tools~Modes~Scree for further details. The option toggles the logarithmic axis transformation, but the plot is redrawn only if the cumulative option is switched off.

Broken Stick

The broken stick can be superimposed on the scree plot. The stick provides one method for determining how many modes should be retained. The stick is shown as a green line. The Broken Stick option toggles the plotting of the broken stick.

Title

The title for any graphic can be modified using the Customise ~ Title menu item available by right-clicking on the graphic. Select the relevant graphic (if prompted) and set a new title using a maximum of 52 characters.
Part IV. CPT Linux Version Tutorial
# Table of Contents

20. Linux Version Introduction ................................................................. 68
21. Linux Version Installation ................................................................. 69
   System and Compiler Requirements ..................................................... 69
   How to Build CPT ............................................................................... 69
   How to Run CPT ............................................................................... 69
Chapter 20. Linux Version Introduction

The CPT Linux source code is a subset of the CPT Windows source code. It contains only the data files IO, LAPACK and core algorithms. Use a Fortran compiler to create the machine dependent executable ‘CPT.x’. The CPT.x will launch an interactive screen with numbered options. Users can interact with the CPT program from the prompt. The Linux version has limited functionality compared to the Windows version. There is no graphic interface so no graphics can ben created, but it can generate output files in CPT V10 format. Users can put all conditions and options in a text file and let CPT run in batch mode to generate all desire output files for one command. This is most useful when it is combined with other program which use CPT output files as input data for other programs. The CPT Linux source code is not downloadable from the public webpage but is freely available on request by sending an e-mail to <cpt@iri.columbia.edu>.
Chapter 21. Linux Version Installation

System and Compiler Requirements

CPT should be able to install on any Linux, Unix or Unix alike system such as Red Hat, Cygwin, and Mac OS. We have tested with gfortran version 4.1.1 or above. Users may be able to compile with other Fortran compilers such NAG f95, pgf90, pgf95 or ifort, etc. but the results are not guaranteed.

How to Build CPT

Type the following commands to build CPT.

```bash
tar -xvzf CPT.10.02.tar.gz
cd CPT/10.02
make all
make deepclean -- clean up top directory and subdirectories
make clean     -- only clean top directory, will not clean subdirectories
```

Copy CPT.x and CPT.ini to your installing directory

You can launch CPT.x from your installing directory. CPT.ini must be in your installed directory.

How to Run CPT

User can run CPT.x interactively through terminal. After typing CPT.x or ./CPT.x, user will see a terminal screen as below.
By typing 1, 2 or 3 and 'return', the user can select the type of analysis they wish to perform. For example, if "2" is entered for PCR, the following screen will appear.

```
Select option:
1. Open X input file
2. Open Y input file
7. Length of training period
8. Length of cross-validation window
EDIT
OPTIONS
35. Change numbers of X EOF modes
36. Change X EOF options
41. Tailoring
42. Transform Y data (turn on)
43. Zero-bound (turn on)
44. Missing value options
45. Bootstrap settings
46. Forecast settings
47. Skill map p-values (turn on)
VIEW
48. Canonical Correlation Analysis (CCA)
50. Multiple Linear Regression (MLR)
0. Exit
```

Again the user can type in the number corresponding to the selected option. The benefit of using the Linux version of CPT is that the user can also save all desired options in a text file then run CPT.x in batch mode. Here is an example text file with all the user selected options. (pcr.txt).
In this example, the user will perform PCR analysis and save some outputs. To run CPT.x in batch mode type the following at the prompt:

```bash
CPT.x < pcr.txt
```

After user types the above command, the CPT.x program will begin printing out text on the screen.
Acknowledgements

CPT was developed using version 3.1.1 of LAPACK for the CCA and PCR algorithms, and Algorithm AS 27 from Applied Statistics 19(1) to calculate upper-tail areas under the Student's t-distribution. The high-resolution country borders, coastlines, lakes and inland seas are extracted from the Environmental Systems Research Institute (ESRI) ArcWorld dataset.

Substantial input on the design of the software was provided by Ousmane Ndiaye, and Theodore Marguerite.

This CPT User Guide was written with the assistance of Lulin Song, Ashley Curtis and Amy Lafferty.
Appendix A. CPT Frequently Asked Questions and Answers

Q: Why does the program sometimes generate windows that do not fit on the screen?

A: The software is designed ideally to operate at a screen resolution of at least 1024 by 768 pixels, but should work at most resolutions of 800 by 600 pixels or finer. Most problems should have been fixed at version 3.04, but it is possible that a window will not format properly at some screen resolutions, and will cause the window to be too wide for the screen. If this happens, please identify your screen resolution, by opening the Control Panel under the Start menu, then selecting the Display icon, and clicking on the Settings tab. The screen resolution should be indicated. Alternatively, you can right click on the desktop, choose properties, and the Display Properties window will open. Then choose the Settings tab to obtain the resolution. Please e-mail your screen resolution to <cpt@iri.columbia.edu> and indicate which window does not fit on the screen. We will try to fix the problem by the next version.

Q: How can I obtain information on the science of climate prediction, and prediction methods, and downscaling methods?


Q: Do you have any plans to develop a Linux version?

A: A source code version of CPT is available on special request. This version could be compiled under Linux and other platforms. The version does not contain any of the graphics features of the Windows version, but does incorporate all of the mathematical functionality of CPT. To request the source code version, please send an email to <cpt@iri.columbia.edu>.

Q: Is it possible to use more than one field of predictors (for example 850 and 500 hPA geopotential heights) in CPT simulataneously?

A: CPT is designed to take only one field of predictors at a time, but it is possible to get the software to produce results using multiple fields. Run the software using one of the predictor fields, and with the number of X EOF modes set to the maximum (this will be the minimum of the number of gridpoints and the length of the training period). Then save the principal component scores, using Data Output. Repeat for the other predictor fields. Now combine the various output files of principal component scores so that the principal components for all the predictor fields are in one file. CPT can then be run with this new file as the predictor variables read in as an unreferenced dataset. Set the X EOF option to “covariance matrix” in order to retain the relative importance of the EOFs. (Note that it may be necessary to weight the two input datasets if they have been derived from different resolution datasets, or have different domains.) Although it will not be possible to view the loadings maps for the combined fields, all the validation results and forecasts will be as if the software had been run with multiple input fields.

Q: Why do I sometimes get an error message when I try to save Project files or save output files?

A: The default installation directory for versions of CPT prior to 4.01 was C:\Program Files\CPT. The project and output files directories are subdirectories of this. Only administrative users have write access to this directory, and so a user without administrator privileges will not be able to save any files in the default directories. An immediate solution is to save the files elsewhere (although this solution will not
resolve the problem of resetting defaults). One more permanent solution is to install CPT in a directory that is not protected, such as C:\CPT, which is the default directory since version 4.01.

**Q:** Why are the forecast probabilities sometimes bimodal; i.e., why is the probability for the normal category sometimes lower than for both the above- and below-normal categories?

**A:** The forecast probabilities are calculated by assuming that the forecast error is normally distributed. This assumption should be reasonable if the Y data are normally distributed, and if they are normally distributed the bimodal forecast probabilities will not occur. However, if the Y data are not normally distributed, bimodal forecast probabilities may occur. Ideally the data should be transformed to a normal distribution before using CPT. There are plans to include a transformation option in the program at a later release.

**Q:** Why does the software sometimes give a message saying "The file contains no non-missing series"?

**A:** If one or both of the input files contains missing values, CPT will attempt to replace these using one of the procedures specified in the Options, Missing Values menu item. However, the missing values are only replaced in those series for which the proportion of missing values is less than a threshold maximum (i.e., CPT will not replace missing values in series where there are too many missing values. This threshold maximum is set by adjusting the "Maximum % of missing values". If this value is too low, then series will be omitted that have only a few values, and if all series are omitted the message above is issued. To avoid this problem, increase the maximum % of missing values. If that still does not work try resetting the length and/or starting date of the training period.

**Q:** When should one use PCR and when CCA?

**A:** The difference between CCA and PCR is fairly small, and, in practice, differences in the results of the two methods are usually minor. In PCR each of the Y variables (predictands) is predicted individually (separate regression models are made for each. PCR is identical to multiple regression except that the predictors are the principal components of the X variables rather than the original X variables. In CCA the Y variables are predicted simultaneously by predicting spatial patterns. In practice, PCR is better suited for predicting either a single predictand, or for a set of predictands that are not expected to have a strong relationship to each other. So if you are making predictions for a single variable or for a few variables that are uncorrelated, use PCR, otherwise it is slightly better to use CCA.

To send comments/questions, go to this online form. http://iri.columbia.edu/outreach/software/cpt/contact_cpt.html

**Q:** How do I register to receive automated update notifications about CPT?

**A:** We have separate mailing lists for Windows and Linux versions of CPT. Register for Windows version updates here and Linux version updates here.