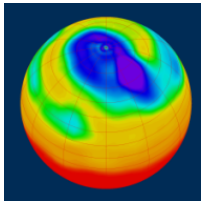
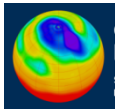
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Reading and writing other formats

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Other formats

- When processing data you will come across a range of formats.
- Common text formats are:
  - XML
  - HTML
  - Json
- Common binary formats in our community are:
  - Grib
  - ESRI shapefiles

Reading XML – Element Tree

```
from urllib.request
import urlopen
from xml.etree.ElementTree import parse
# Download the RSS feed and parse it
u = urlopen('http://planet.python.org/rss20.xml')
doc = parse(u)
# Extract and output tags of interest
for item in doc.iterfind('channel/item'):
    title = item.findtext('title')
    date = item.findtext('pubDate')
    link = item.findtext('link')

print title, date, link
```

Reading HTML - HTMLParser

```
from HTMLParser import HTMLParser

# create a subclass and override the handler methods
class MyHTMLParser(HTMLParser):
    def handle_starttag(self, tag, attrs):
        print "Encountered a start tag:", tag
    def handle_endtag(self, tag):
        print "Encountered an end tag :", tag
    def handle_data(self, data):
        print "Encountered some data  :", data

# instantiate the parser and fed it some HTML
parser = MyHTMLParser()
parser.feed('<html><head><title>Test</title></head>'
            '<body><h1>Parse me!</h1></body></html>')
```

Reading Grib

- See later section on Iris

Reading ESRI Shapefiles

```
import shapefile
sf = shapefile.Reader("shapefiles/blockgroups.shp")
shapes = sf.shapes()

# Get the bounding box of the 4th shape.
# Round coordinates to 3 decimal places
bbox = shapes[3].bbox
print ['%.3f' % coord for coord in bbox]
```

```
['-122.486', '37.787', '-122.446', '37.811']
```