O quarto paradigma: a ciência da análise de dados

Bruno Ferrero[1] Gabriel Pato[1]

[1]Instituto de Matemática e Estatística

Universidade de São Paulo

23 de Novembro de 2011

Roteiro

O 4° Paradigma

Dados e as ciências climáticos

Dados e as ciências biológicas

Conclusão

O quarto paradigma: e-Science

blog *The Fourth Paradigm*

- ciência e quebra de paradigmas;
- teoricos;
- ciência empírica;
- simulações;
- e-Science;

O quarto paradigma: e-Science

blog *The Fourth Paradigm*

- ciência e quebra de paradigmas;
- teoricos;
- ciência empírica;
- simulações;
- e-Science;

O quarto paradigma: e-Science

blog *The Fourth Paradigm*

- ciência e quebra de paradigmas;
- teoricos;
- ciência empírica;
- simulações;
- e-Science;

O quarto paradigma: e-Science

blog *The Fourth Paradigm*

- ciência e quebra de paradigmas;
- teoricos;
- ciência empírica;
- simulações;
- e-Science;

O quarto paradigma: e-Science

e-Science

- uso intensivo de tecnicas computacionais;
  - ex. novas tecnicas de visualização de dados;
- grande volume de dados ;
- grupos de pesquisadores (diversos coloaboradores);
- ferramenta X paradigma;

Dados e o clima

- estudar o clima $\rightarrow$ analisar dado;
- duas fontes de dados:
  - observacional;
  - dados numéricos;
- dado é caro!!;

Dados climáticos

- estudar o clima $\rightarrow$ analisar dado;
- duas fontes de dados:
  - observacional;
  - dados numéricos;
- dado climático observado é caro!!;
- avanço da computação $\rightarrow$ mais dados numéricos;

Dados climáticos

- estudar o clima $\rightarrow$ analisar dado;
- duas fontes de dados:
  - observacional;
  - dados numéricos;
- dado climático observado é caro!!;
- avanço da computação $\rightarrow$ mais dados numéricos;

Dados climáticos

- estudar o clima $\rightarrow$ analisar dado;
- duas fontes de dados:
  - observacional;
  - dados numéricos;
- dado climático observado é caro!!;
- avanço da computação $\rightarrow$ mais dados numéricos;

Dados climáticos

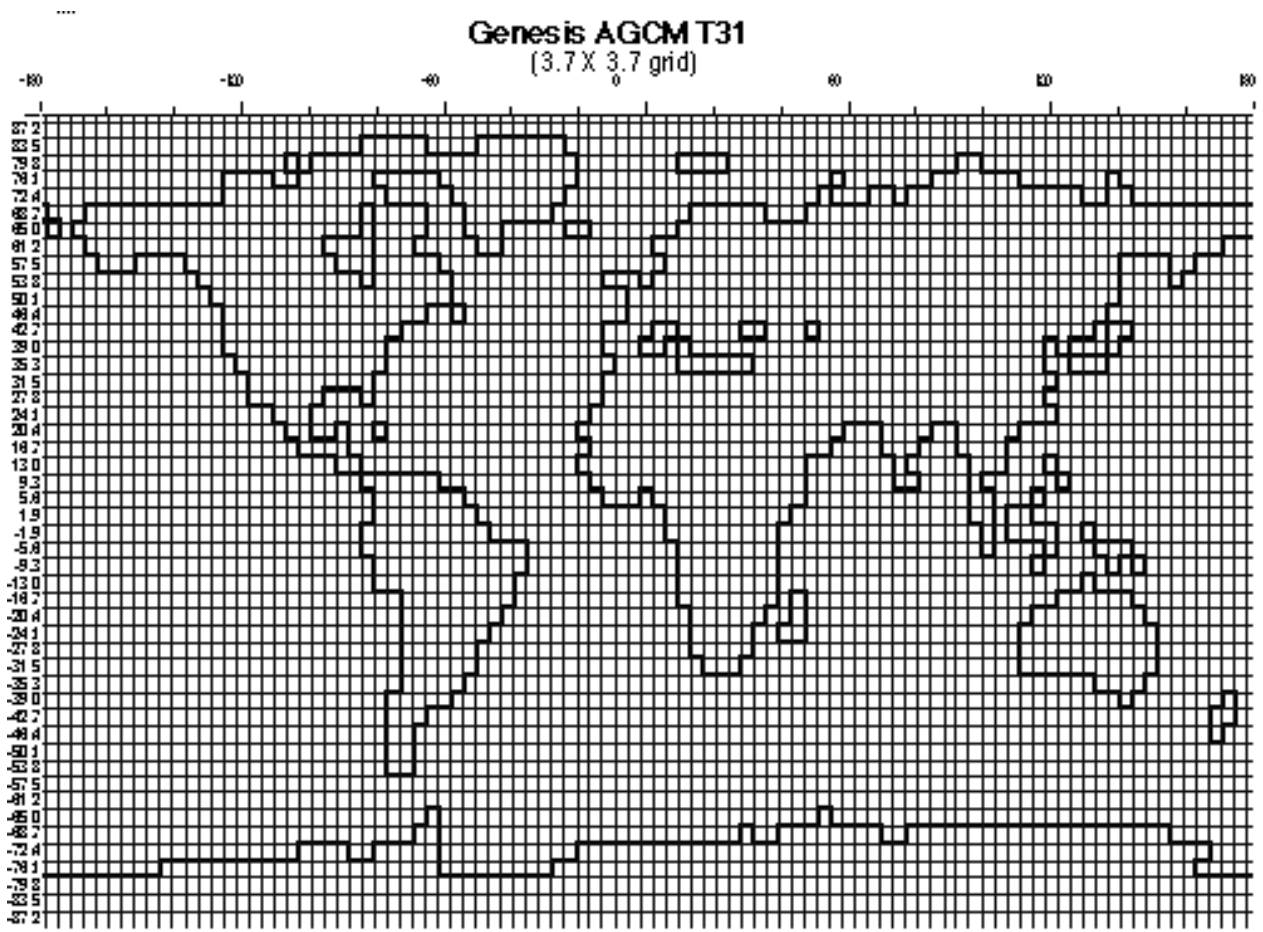
- estudar o clima $\rightarrow$ analisar dado;
- duas fontes de dados:
  - observacional;
  - dados numéricos;
- dado climático observado é caro!!;
- avanço da computação $\rightarrow$ mais dados numéricos;

Simulação numérica do clima

- reproduzir a dinâmica do clima por meio de modelos matemáticos;
- alto poder de processamento;
- alta capacidade de armazenamento;
- super computadores, ex. *Earth Simulator*;
- mais processamento $\rightarrow$ modelos mais complexos $\rightarrow$ simulação mais próxima da realidade;

Simulação numérica do clima

gera um enorme volume de dados



- 110 variáveis climáticas;
- 360 x 180 x 30 resolução espacial (x, y, z);
- 6 horas frequência temporal

Simulando um perído de 100 anos:

$$110 \times (4 \times 365 \times 100) \times (360 \times 180 \times 30) = \text{Muito dado!!}$$

- 110 variáveis climáticas;
- 360 x 180 x 30 resolução espacial (x, y, z);
- 6 horas frequência temporal

Simulando um perído de 100 anos:

$$110 \times (4 \times 365 \times 100) \times (360 \times 180 \times 30) = \text{Muito dado!!}$$

- 110 variáveis climáticas;
- 360 x 180 x 30 resolução espacial (x, y, z);
- 6 horas frequência temporal

Simulando um perído de 100 anos:

$$110 \times (4 \times 365 \times 100) \times (360 \times 180 \times 30) = \text{Muito dado!!}$$

- 110 variáveis climáticas;
- 360 x 180 x 30 resolução espacial (x, y, z);
- 6 horas frequência temporal

Simulando um perído de 100 anos:

$$110 \times (4 \times 365 \times 100) \times (360 \times 180 \times 30) = \text{Muito dado!!}$$

- 110 variáveis climáticas;
- 360 x 180 x 30 resolução espacial (x, y, z);
- 6 horas frequência temporal

Simulando um perído de 100 anos:

$$110 \times (4 \times 365 \times 100) \times (360 \times 180 \times 30) = \text{Muito dado!!}$$

Organização de dados

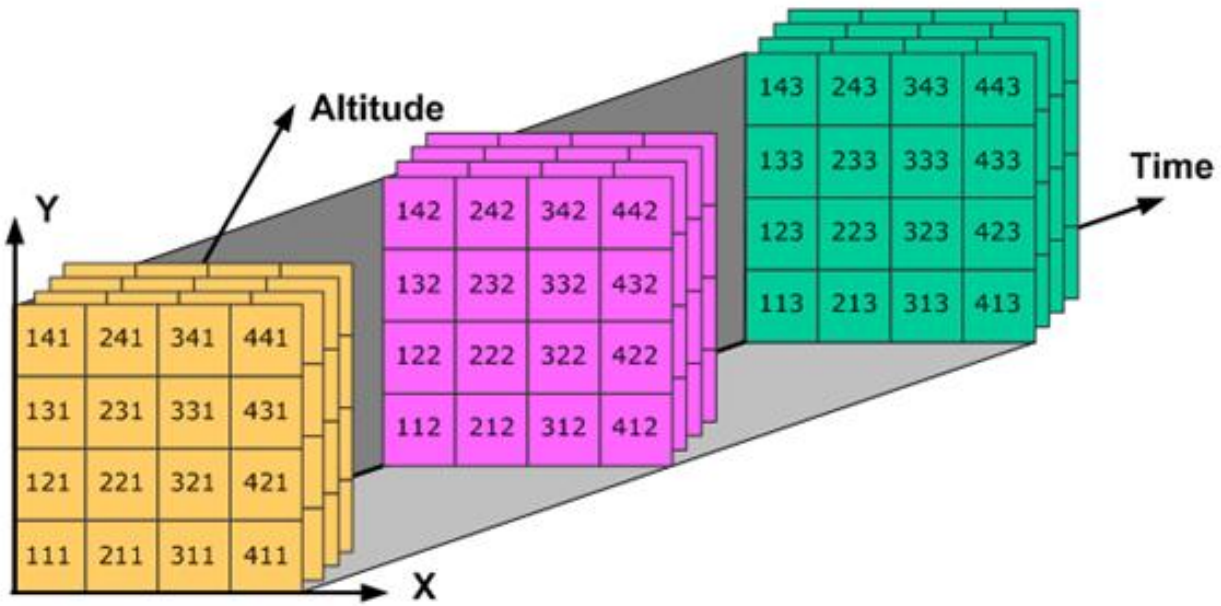
Características

- portabilidade;
- auto descritivo;
- padronização;

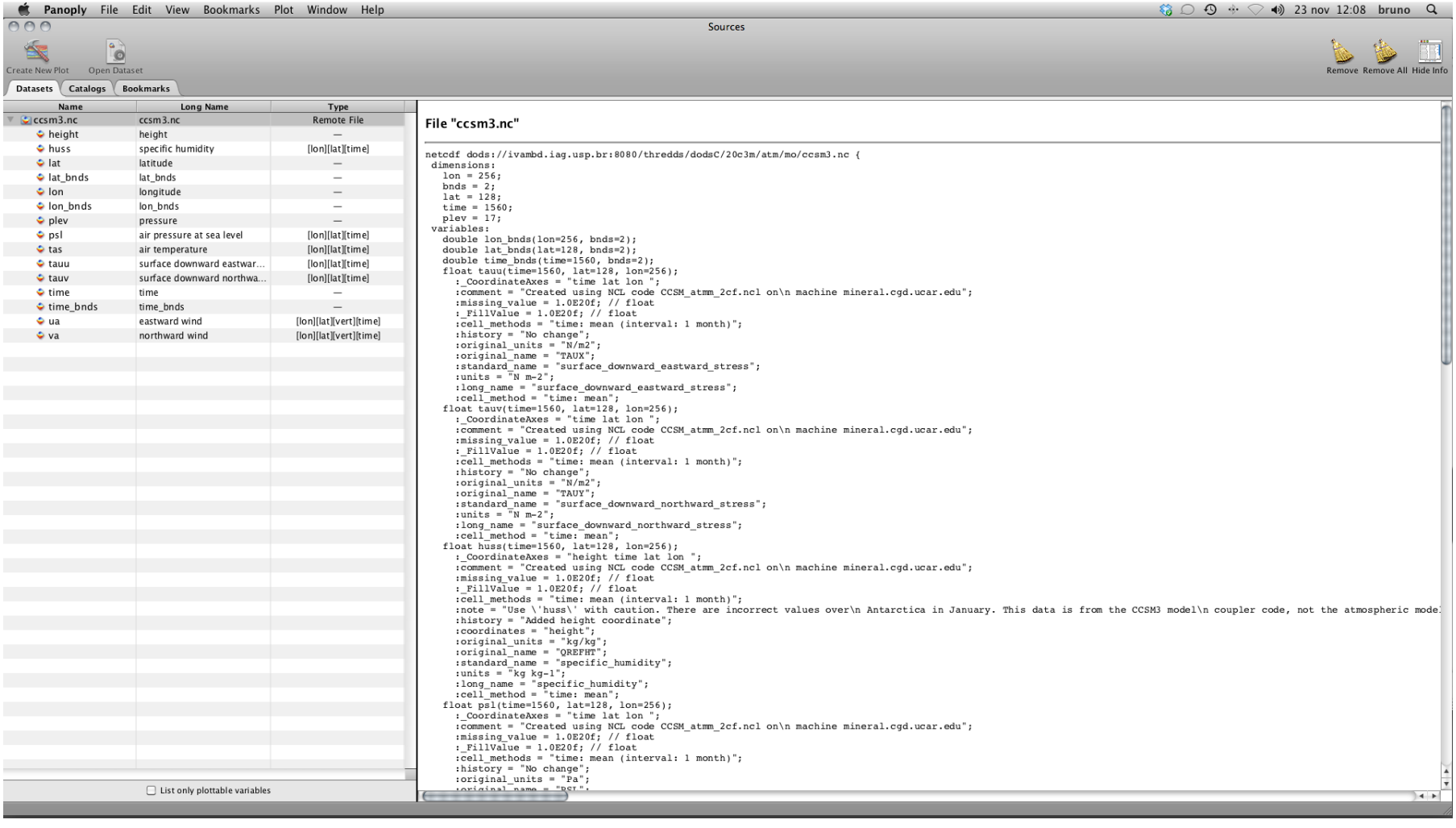
netCDF

- *software livre*
- auto descritivo;
- portabilidade
- binário com um cabeçalho ASCII;
- convenção;
- *gridded data*
- trabalha com múltiplas dimensões, $(x, y, z, time)$;

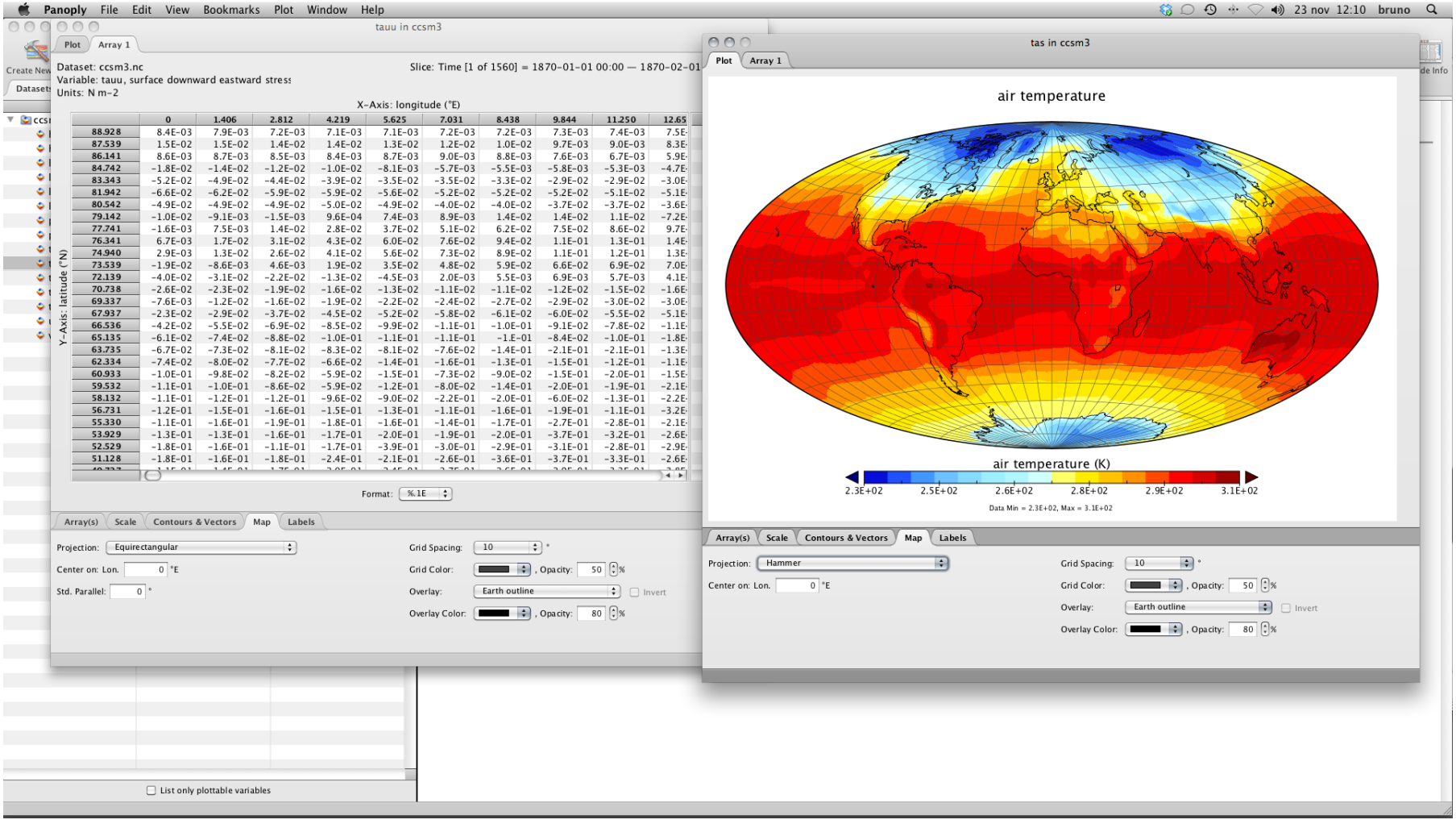
Gridded data



Exemplo


Panoply File Edit View Bookmarks Plot Window Help
23 nov 12:08 bruno
Sources
Create New Plot
Open Dataset
Remove
Remove All
Hide Info
Datasets
Catalogs
Bookmarks
Name
Long Name
Type
ccsm3.nc
ccsm3.nc
Remote File
height
height
—
huss
specific humidity
[lon][lat][time]
lat
latitude
—
lat_bnds
lat_bnds
—
lon
longitude
—
lon_bnds
lon_bnds
—
plev
pressure
—
psl
air pressure at sea level
[lon][lat][time]
tas
air temperature
[lon][lat][time]
tauu
surface downward eastwar...
[lon][lat][time]
tauv
surface downward northwa...
[lon][lat][time]
time
time
—
time_bnds
time_bnds
—
ua
eastward wind
[lon][lat][vert][time]
va
northward wind
[lon][lat][vert][time]
List only plottable variables
File "ccsm3.nc"
netcdf dods://ivambd.iag.usp.br:8080/thredds/dodsC/20c3m/atm/mo/ccsm3.nc {
dimensions:
lon = 256;
bnds = 2;
lat = 128;
time = 1560;
plev = 17;
variables:
double lon_bnds(lon=256, bnds=2);
double lat_bnds(lat=128, bnds=2);
double time_bnds(time=1560, bnds=2);
float tauu(time=1560, lat=128, lon=256);
:_CoordinateAxes = "time lat lon ";
:comment = "Created using NCL code CCSM_atmm_2cf.ncl on\n machine mineral.cgd.ucar.edu";
:missing_value = 1.0E20f; // float
:_FillValue = 1.0E20f; // float
:cell_methods = "time: mean (interval: 1 month)";
:history = "No change";
:original_units = "N/m2";
:original_name = "TAUX";
:standard_name = "surface_downward_eastward_stress";
:units = "N m-2";
:long_name = "surface_downward_eastward_stress";
:cell_method = "time: mean";
float tauv(time=1560, lat=128, lon=256);
:_CoordinateAxes = "time lat lon ";
:comment = "Created using NCL code CCSM_atmm_2cf.ncl on\n machine mineral.cgd.ucar.edu";
:missing_value = 1.0E20f; // float
:_FillValue = 1.0E20f; // float
:cell_methods = "time: mean (interval: 1 month)";
:history = "No change";
:original_units = "N/m2";
:original_name = "TAUY";
:standard_name = "surface_downward_northward_stress";
:units = "N m-2";
:long_name = "surface_downward_northward_stress";
:cell_method = "time: mean";
float huss(time=1560, lat=128, lon=256);
:_CoordinateAxes = "height time lat lon ";
:comment = "Created using NCL code CCSM_atmm_2cf.ncl on\n machine mineral.cgd.ucar.edu";
:missing_value = 1.0E20f; // float
:_FillValue = 1.0E20f; // float
:cell_methods = "time: mean (interval: 1 month)";
:note = "Use \'huss\' with caution. There are incorrect values over\n Antarctica in January. This data is from the CCSM3 model\n coupler code, not the atmospheric mode
:history = "Added height coordinate";
:coordinates = "height";
:original_units = "kg/kg";
:original_name = "QREFHT";
:standard_name = "specific_humidity";
:units = "kg kg-1";
:long_name = "specific_humidity";
:cell_method = "time: mean";
float psl(time=1560, lat=128, lon=256);
:_CoordinateAxes = "time lat lon ";
:comment = "Created using NCL code CCSM_atmm_2cf.ncl on\n machine mineral.cgd.ucar.edu";
:missing_value = 1.0E20f; // float
:_FillValue = 1.0E20f; // float
:cell_methods = "time: mean (interval: 1 month)";
:history = "No change";
:original_units = "Pa";

Exemplo


Panoply File Edit View Bookmarks Plot Window Help
23 nov 12:10 bruno
tauu in ccsm3
Plot
Array 1
Dataset: ccsm3.nc
Variable: tauu, surface downward eastward stress
Units: N m-2
Slice: Time [1 of 1560] = 1870-01-01 00:00 — 1870-02-01
X-Axis: longitude (°E)
Y-Axis: latitude (°N)
Format:
Array(s)
Scale
Contours & Vectors
Map
Labels
Projection: Equirectangular
Center on: Lon.
Std. Parallel:
Grid Spacing:
Grid Color:
Overlay: Earth outline
Overlay Color:
Opacity:
List only plottable variables
tas in ccsm3
air temperature
air temperature (K)
2.3E+02
2.5E+02
2.6E+02
2.8E+02
2.9E+02
3.1E+02
Data Min = 2.3E+02, Max = 3.1E+02
Projection: Hammer

Painel Intergovernamental para as Mundanças

Climáticas - IPCC

- cenários de desenvolvimento para fazer projeções climáticas;
- uso intensivo de modelos climáticos;
- milhares de pesquisadores em todo o mundo;
- avaliação colaborativa das simulações;

Painel Intergovernamental para as Mundanças Climáticas - IPCC

- cenários de desenvolvimento para fazer projeções climáticas;
- uso intensivo de modelos climáticos;
- milhares de pesquisadores em todo o mundo;
- avaliação colaborativa das simulações;

Dados e as ciências biológicas

...

- Genoma;
- Proteoma;
- Metaboloma;

Genoma e Proteoma

Tipos de dados gerados

- sequência de nucleotídeos;
- sequência de aminoácidos;
- dobramento de proteínas;

Genoma

- Sequenciamento genético;
  - sequenciadores:
    - Pyrosequencing $\rightarrow$ 40 milhões de pares de bases por hora;
  - banco de dados:
    - GenBank
- necessidade de tratamento/análise;

GenBank

Características

- criação: 1982
- acesso livre
- atualmente, mais de 140 milhões de sequências

Formatos de arquivos

- ASN.1
- GenBank
- outros ...


```
FEATURES             Location/Qualifiers
     source          1..428
                     /organism="Macaca mulatta"
                     /mol_type="mRNA"
                     /strain="Indian"
                     /db_xref="taxon:9544"
                     /clone="IBIUW:32275"
                     /sex="female"
                     /dev_stage="adult"
                     /lab_host="Electromax DH10B"
                     /clone_lib="Katze_MMOV"
                     /note="Organ: ovary; Vector: pDONR 222; Site_1: BsrG I;
                     Site_2: BsrG I; Created from CloneMiner cDNA Library
                     Construction kit (catalog #18249-029)"
ORIGIN
        1 ttggctcttc tacctgcaac cgaatgcttg atgaagccac cagtgccctg acagaggagg
       61 tggagaatga gctctatcgc atcggccagc agctggggat gacgttcatc agtgtgggac
      121 atcggcagag ccttgagaag tttcattcct tcgttctgaa actctgtgga ggaggaagat
      181 gggagctgat gagaatcaaa gtggaatgaa gctccagctt ttagaaggag agccacactc
      241 tggagggtcg gcagccctca ggagtgacca ggaggactgg cggggaagat cgagctcagg
      301 ttcgccacat aggtcctgtg caggagccct ggcggtgttg ggctgagccc gggtctggat
      361 ttctgtgggg gacactgagt ctcccagtgt tcagtctccc aggactctgc tgcctcagcc
      421 agagcctc
```

Conclusões

- importância da organização de dados;
- ambiente de compartilhameno;
- levando a um sistema de produção colaborativa.

Obrigado!