

GRIB ↔ NetCDF

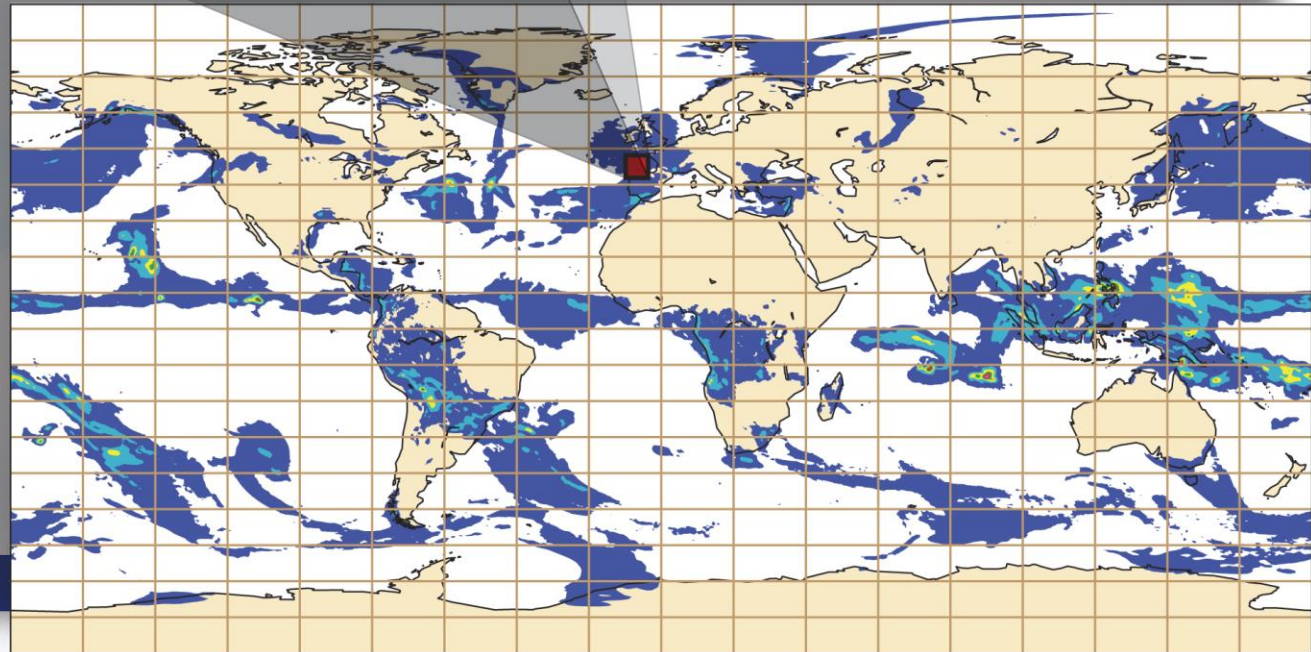
Setting the scene



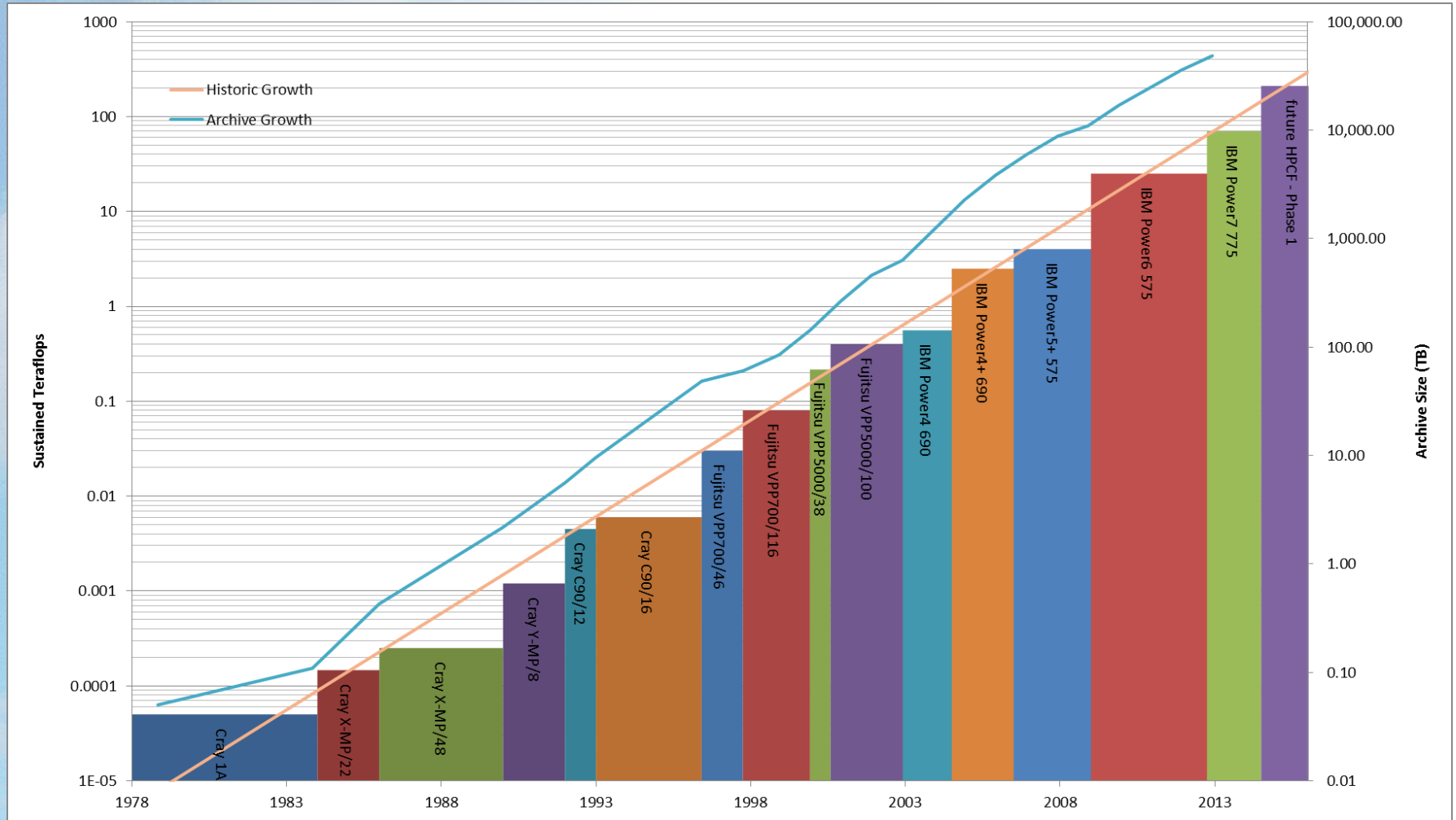
ECMWF model output are encoded in GRIB

- One parameter
- One date
- One time
- One step
- One level
- One forecasting system
- ...

7.2	9.9	3.6	0.4	8.3	0.2	0.5	0.1	9.1	6.7
0.3	8.8	1.8	0.5	0.3	0.1	2.7	0.1	7.9	6.9
7.1	9.2	3.6	0.4	8.3	0.2	6.5	3.3	5.5	5.3
2.2	1.1	1.7	0.7	3.5	2.4	0.8	1.9	9.0	6.7
5.1	0.9	1.9	8.9	5.9	0.4	1.5	2.0	7.7	0.7
6.2	0.4	1.4	9.8	9.9	7.7	0.9	3.2	7.2	4.8
8.1	1.4	4.4	0.4	0.3	7.2	3.5	3.4	1.1	9.7
7.0	3.6	4.9	0.7	6.8	1.2	0.1	2.2	6.6	6.0
0.2	7.7	3.6	3.1	8.6	0.5	9.5	0.8	5.6	5.0
3.2	7.2	3.1	0.4	0.9	0.3	0.7	0.4	0.2	0.0



Size of the archive vs. Sustained HPC performance



GRIB are stored in MARS

- 28 years in existence
- A managed archive
- MARS is not a file system
 - Users are not aware of the location of the data
 - Retrievals are expressed in meteorological terms
- An archive, not a database
 - Metadata online
 - Data offline (automated tape library)



MARS in numbers

- 53 Petabytes of primary data in ~ 11 million files, for more than 170 billion ($1.7 \cdot 10^{11}$) meteorological fields
- ~ 800 Gigabytes of metadata
- 200 million fields added daily (peaks at 100 Terabytes)
- 650 active users/day executing 1.5 million requests/day
- ~ 100 Terabytes retrieved daily

What data?

- Operational runs
 - Medium-range (15 days, twice a day, including ensemble)
 - Extended-range (a month), Long-range (a year)
 - Re-forecasts , Ocean waves
- Projects
 - Reanalyses (15 years, 45 years, 100 years)
 - WMO: TIGGE, TIGGE-LAM, S2S
 - EU projects: DEMETER, ENSEMBLES, EURO4M, MACC, PROVOST, ECSN...
- Research experiments
 - ECMWF, Member States
- Member States' Projects
 - COSMO-LEPS , Aladin-LEAF

MARS language – Retrieve request

```
retrieve,  
  class      = od,  
  stream     = oper,  
  expver     = 1,  
  
  date       = -3,  
  time       = 12,  
  
  type       = analysis,  
  levtype    = model levels,  
  levelist   = 1/to/137,  
  param      = temperature,  
  
  grid       = 2.5/2.5,  
  target     = "analysis"
```

action
 identification

 date & time related

 data related

 post-processing
 storage

Metview and GRIB

Metview - GRIB Examiner < @anilib>

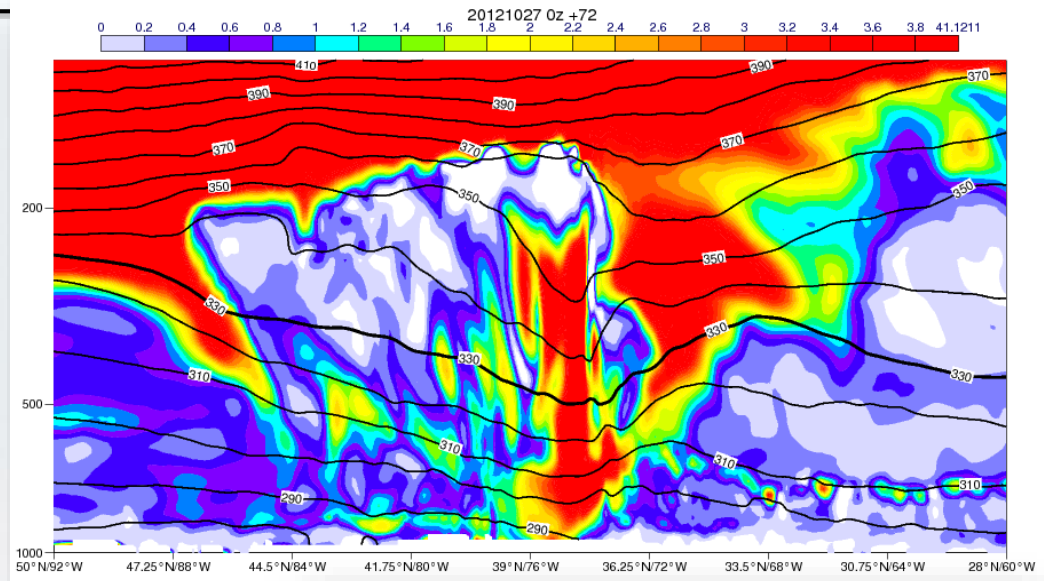
File Edit View Profiles Help

Key profile: /nv System: Default

File: /mnt/psfmp/dp/ig/ig/7342/mv4436/cgi/mars3P/UPD
 Parameters: /mnt/psfmp/dp/ig/ig/7342/mv4436/cgi/mars3P/UPD
 Total number of messages: 34

Index	Name	Date	Time	Step	Level	LevTyp
31	u	2011021	0000	8	850	pl
32	v	2011021	0000	8	850	pl
33	u	2011021	0000	8	850	pl
34	v	2011021	0000	8	850	pl
35	u	2011021	0000	12	850	pl
36	v	2011021	0000	12	850	pl
37	u	2011021	0000	18	850	pl
38	v	2011021	0000	18	850	pl
39	u	2011021	0000	24	850	pl
40	v	2011021	0000	24	850	pl
41	u	2011021	0000	30	850	pl
42	v	2011021	0000	30	850	pl
43	u	2011021	0000	36	850	pl
44	v	2011021	0000	36	850	pl
45	u	2011021	0000	42	850	pl
46	v	2011021	0000	42	850	pl
47	u	2011021	0000	48	850	pl
48	v	2011021	0000	48	850	pl
49	u	2011021	0000	54	850	pl
50	v	2011021	0000	54	850	pl
51	u	2011021	0000	60	850	pl
52	v	2011021	0000	60	850	pl
53	u	2011021	0000	66	850	pl
54	v	2011021	0000	66	850	pl
55	u	2011021	0000	72	850	pl
56	v	2011021	0000	72	850	pl
57	u	2011021	0000	78	850	pl
58	v	2011021	0000	78	850	pl
59	u	2011021	0000	84	850	pl
60	v	2011021	0000	84	850	pl

Task: Generating WMO style dump for message: 1
 Command: /mnt/psfmp/dp/ig/ig/7342/mv4436/cgi/mars3P/UPD -s=10-64/mvgrib_dump -o=cour1 -
 /mnt/psfmp/dp/ig/ig/7342/mv4436/cgi/mars3P/UPD
 Status: OK



Metview - uPlot

File View Animation Zoom Tools Help

Thursday 27 October 2011 00 UTC ECMWF Forecast VTSunday 30 October 2011 00 UTC 500 hPa Geopotential
 Thursday 27 October 2011 00 UTC ECMWF Forecast VTSunday 30 October 2011 00 UTC surface 2 metre temperature

Layer	Value	Lon	Lat	Dist (km)
T-fc	0.083866	31.50	67.50	53.55
Z-fc	541.84	31.50	67.50	53.55

statistics - /home/graphics/cgi/metview/Tests/Macros/statistics

File Edit View Insert Program Settings

```
# retrieve some data

f1 = retrieve (date : -1, levels : 1000, grid : [1.5, 1.5])
f2 = retrieve (date : -2, levels : 1000, grid : [1.5, 1.5])

# perform some calculations for comparison

cv_f1f2 = covar_a (f1, f2)
cv_f1f1 = covar_a (f1, f1)
cv_f2f2 = covar_a (f2, f2)
var_f1 = var_a (f1)
var_f2 = var_a (f2)

corr_manual = cv_f1f2 / (sqrt(cv_f1f1) * sqrt(cv_f2f2))
corr_manual2 = cv_f1f2 / (sqrt(var_f1) * sqrt(var_f2))
corr_builtin = corr_a (f1, f2)
```

Choosing RETRIEVE (MARS)
 covar of f1 and f2 = 707195.562425
 corr_manual = 0.876684930973
 corr_manual2 = 0.876684930973
 corr_builtin = 0.876684930973

Program finished (OK) : 4.078 s [Finished at 14:05:55] L: 14, C: 27

Metview and NetCDF

File: /home/graphics/cgi/metview/Tests/uplot/rh850.nc
Permissions: -r-xr-x--- Owner: cgi Group: graphics Size: 520KB Modified: 2008-10-30 09:04

Meta data Ncdump

Parameters	Values
Variables	
longitude	
Type	float
Dimensions	(longitude)
Attributes	
units	degrees_east
long_name	longitude
Data values	
latitude	
time	
r	
Type	short
Dimensions	(time, latitude, longitude)
Attributes	
Data values	

```
2  
3 # Read netcdf file  
4 file_nc = read(nc_filename)  
5  
6  
7 # Find the indexes of the 'special' variables we  
8 vars = variables(file_nc)  
9 time_var_index = find(vars, 'time')  
10 region_var_index = find(vars, 'geo_region')  
11  
12 # extract the times - these are the same for all  
13 setcurrent(file_nc, time_var_index)  
14 times = values(file_nc)  
15 base_time = parse_base_time_attribute(file_nc)  
16 new_times = tolist(times)  
17 new_times = base_time.base + new_times  
18
```

NO TITLE

160°W 140°W 120°W 100°W 80°W 60°W 40°W 20°W 0°E 20°E 40°E

nviz - /Tests/uplot - Metview

Icon name: nviz
Folder: /Tests/uplot
Type: NETCDF_VISUALISER Modified: 2014-09-22 14:07

Netcdf Plot Type	Geo Matrix
Netcdf Filename	OFF
Netcdf Data	rh850.nc
Netcdf Latitude Variable	latitude
Netcdf Longitude Variable	longitude
Netcdf X Variable	

Templates Save OK Cancel

Frames Layers Data

Titles <Coastlines> nviz

Metadata Values

units %

Statistics (for data in visible area)

Points	259920
Minimum	-3.20073
Maximum	107.485
Average	71.0134
Stdev	25.3264
Skewness	-0.820627
Kurtosis	-0.333385

Histogram (for data in visible area)

Bar	From	To	Count
1	-3.20073	0	39

Bar From To Count

But we want more...

We want to archive NetCDF in MARS

...and provide a service on par with what we do
for GRIBs.

GRIB ↔ NetCDF

Setting the scene (part 2)

Disclaimer

GRIB vs. NetCDF: I have no preferences

I know GRIB better, that's all.

GRIB

- Designed for telecoms
 - As small as possible, table driven (must to read the doc :-)
 - Record/message format, in memory
 - No separation between format and data model
 - Used in operational NWP, exchanged on the GTS
 - Designed by committee

 - No such thing as a “GRIB file”, just a file with GRIB messages
- ```
% cat file1.grib file2.grib > file3.grib
```
- `file3.grib` is a valid “GRIB file”



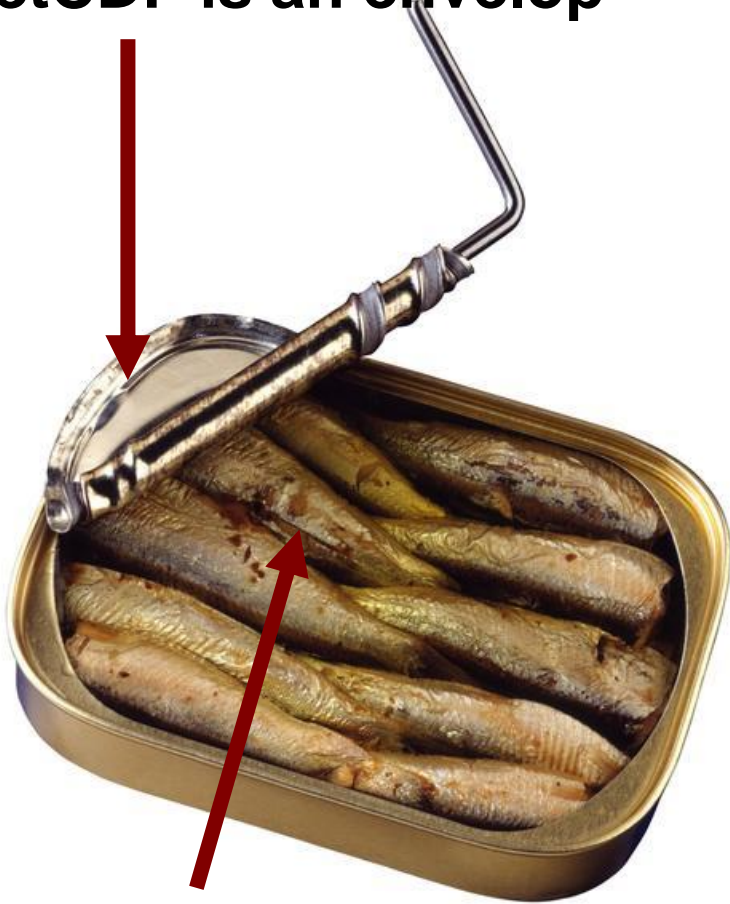
- Primarily an API/library
- Self describing
- Needs a convention (CF)
- Clean separation between format and data model (CF)
- Loads of tools
- Used in academia, oceanography and climate modeling
- CF: community driven
  
- NetCDF is a file format:

```
% cat file1.nc file2.nc > file3.nc
```

- `file3.nc` is NOT a valid NetCDF file.

# NetCDF vs. GRIB/BUFR

NetCDF is an envelop



CF: data stored using a convention



GRIB/BUFR: data and envelop are mixed

# Converting from GRIB to NetCDF

---

- Metadata
  - Date and time
  - Parameters
- Data
  - Units
  - Grids
- Compression
  - Internal (Packing)
  - External (zlib)
- File structures

# Metadata

# Why do we need metadata?

---

- Type 1: we cannot use the data otherwise
  - E.g. description of the grid (latitudes, longitudes)
  - E.g. units
- Type 2: identification (used for indexing, use for querying)
  - E.g. date/time
- Type 3: Nice to have
  - E.g. contact details of principal investigator
- GRIB => NetCDF
  - How to map this metadata?
  - How to map the data?

# Convention? What convention?

---

- Parameter names are well covered by CF.
- What about other attributes:
  - lat/lon?
  - lat/long?
  - latitude/longitude?
  - x/y?
  - y/x?
- And:
  - lev?
  - level?
  - height?
  - z?
- I have seen them all. Are users supposed to inspect new files before using them?

# Units

# Should coersion modify the data?

---

- Example: Total Precipitations
  - NetCDF:  $\text{kg m}^{-2}$  (e.g. mm assuming 1l of water = 1km)?
  - GRIB:  $\text{m m}^{-2}$  ?
  - Mapping implies multiplication by 1000
- Is that acceptable?

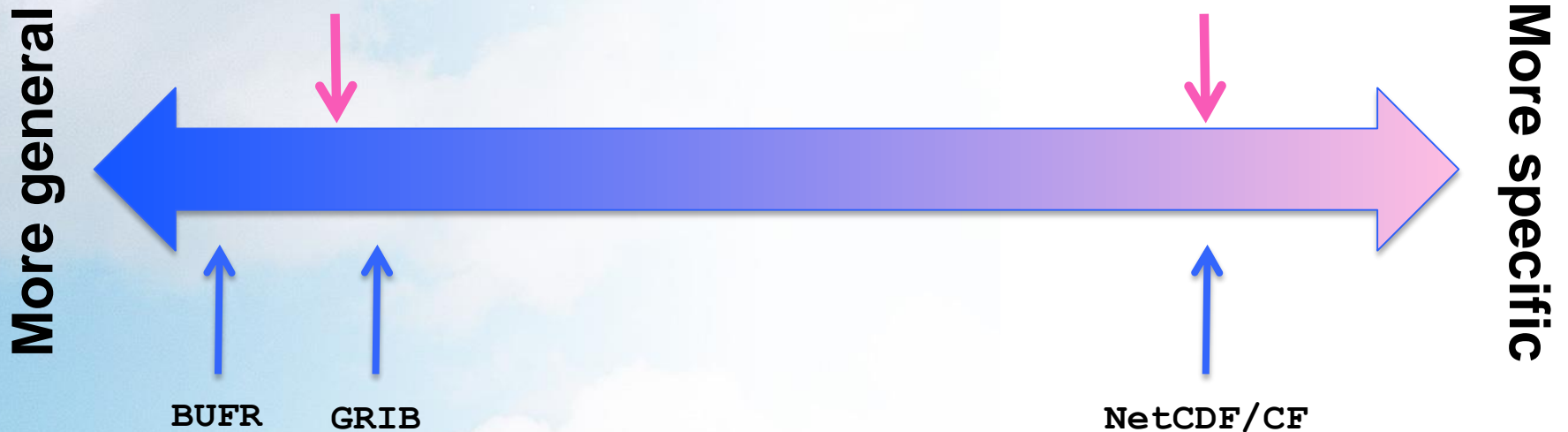


# Parameter names

# “Semantic Spectrum”

```
discipline = meteorology
parameter = temperature
level = 2
level_unit = meter
```

surface\_air\_temperature



*Computer “friendly”*

*Human friendly*

## Two interesting examples:

tendency\_of\_atmosphere\_mass\_content\_of\_  
particulate\_organic\_matter\_dry\_aerosol\_  
expressed\_as\_carbon\_due\_to\_emission\_from\_  
residential\_and\_commercial\_combustion

surface\_upward\_mass\_flux\_of\_carbon\_dioxide\_  
expressed\_as\_carbon\_due\_to\_emission\_from\_  
fires\_excluding\_anthropogenic\_land\_use\_change

# File structure

# File structures

---

- How to structure NetCDF files?
- ECMWF golden rule:
  - File must be self describing
  - File name **MUST NOT** carry any semantic

- Consider:

```
% mv ECMWF-ERA20C-geopotential-20010101.nc foo.nc
```

- What is in foo.nc ?
  - `ncdump` (or `grib_dump`) should tell us

# So how do we structure a NetCDF file?

---

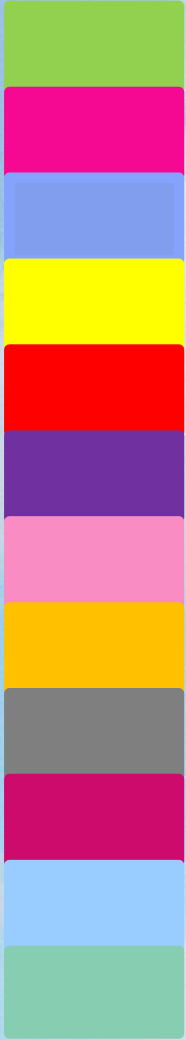
- One 2D field per file?
- One 3D field per file?
- Many 3D fields per file?
- Many 4D files per files?

=> See effect on packing

# File structure: the problem: how to map to NetCDF files?

---

GRIB file



# File structure: one file per field

GRIB file



file1.nc



file2.nc



file3.nc



file4.nc



file5.nc



file6.nc



file7.nc



file8.nc



file10.nc



file9.nc



file11.nc

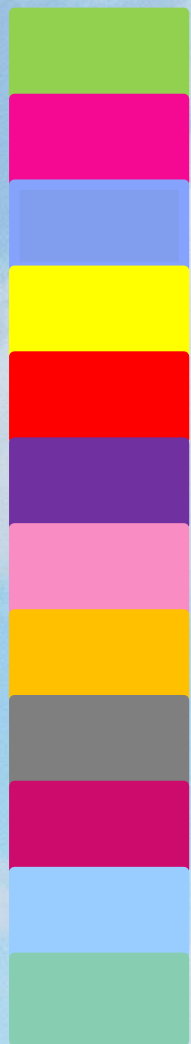


file12.nc



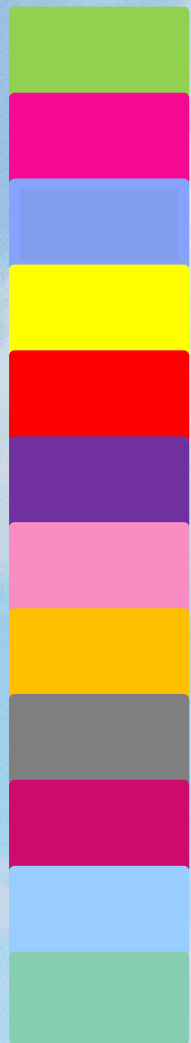
# File structure: group by time? By level? Both?

GRIB file

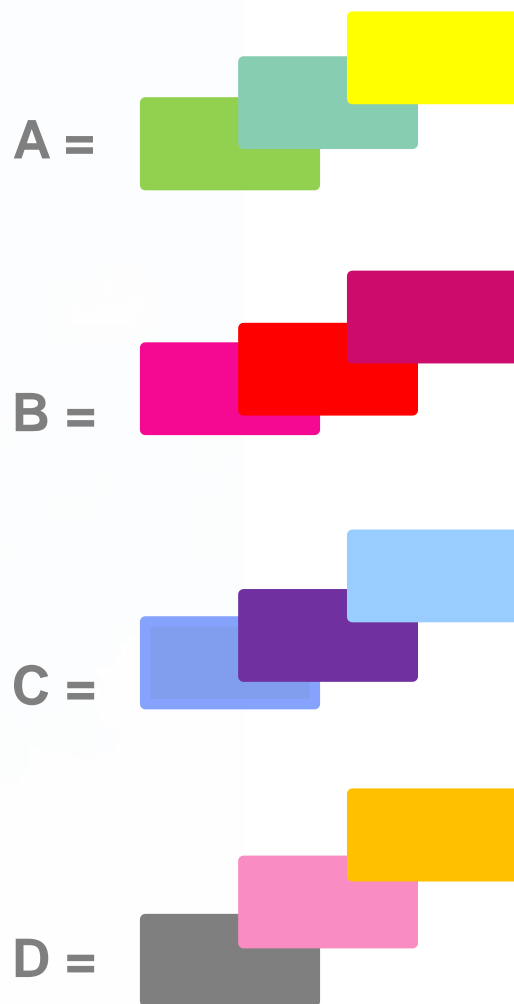


# File structure: group by time? By level? Both?

GRIB file

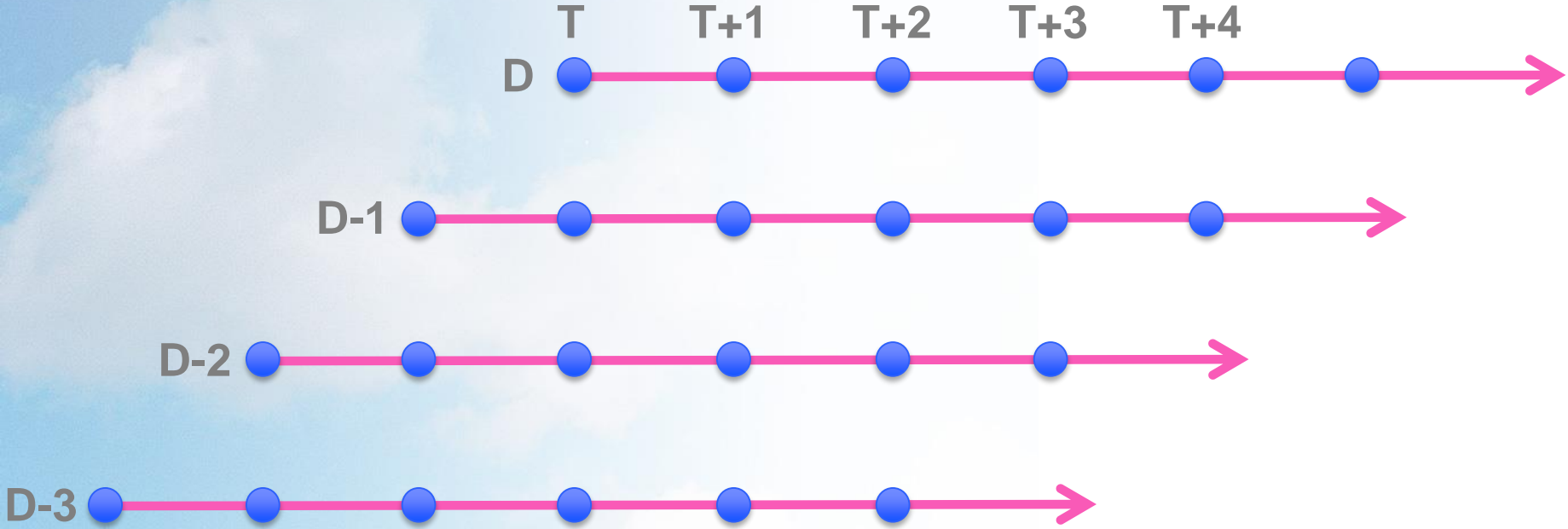


NetCDF file

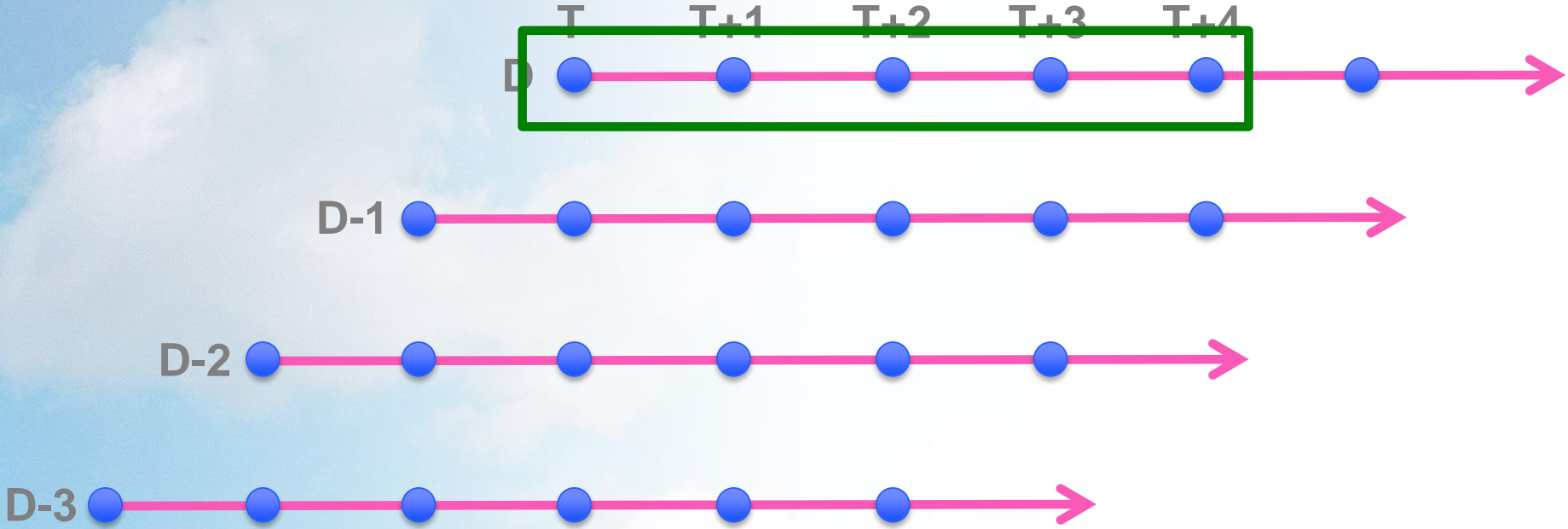


# Date and time

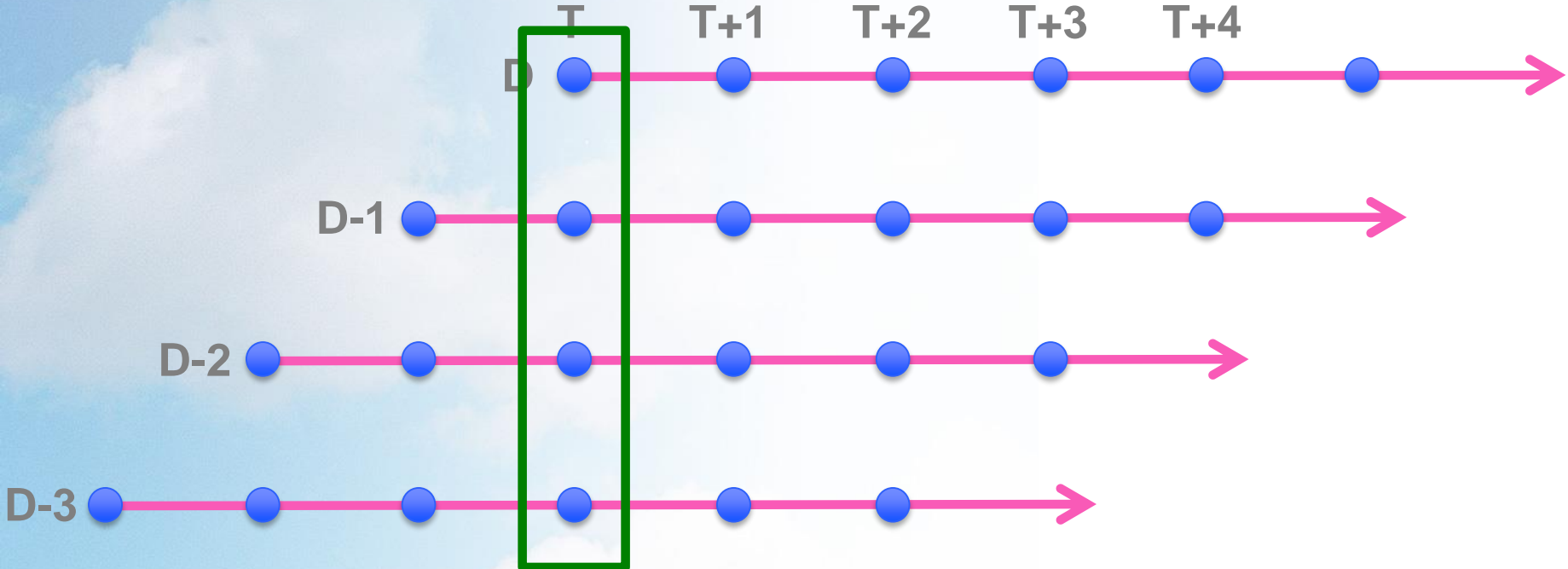
# Date & time



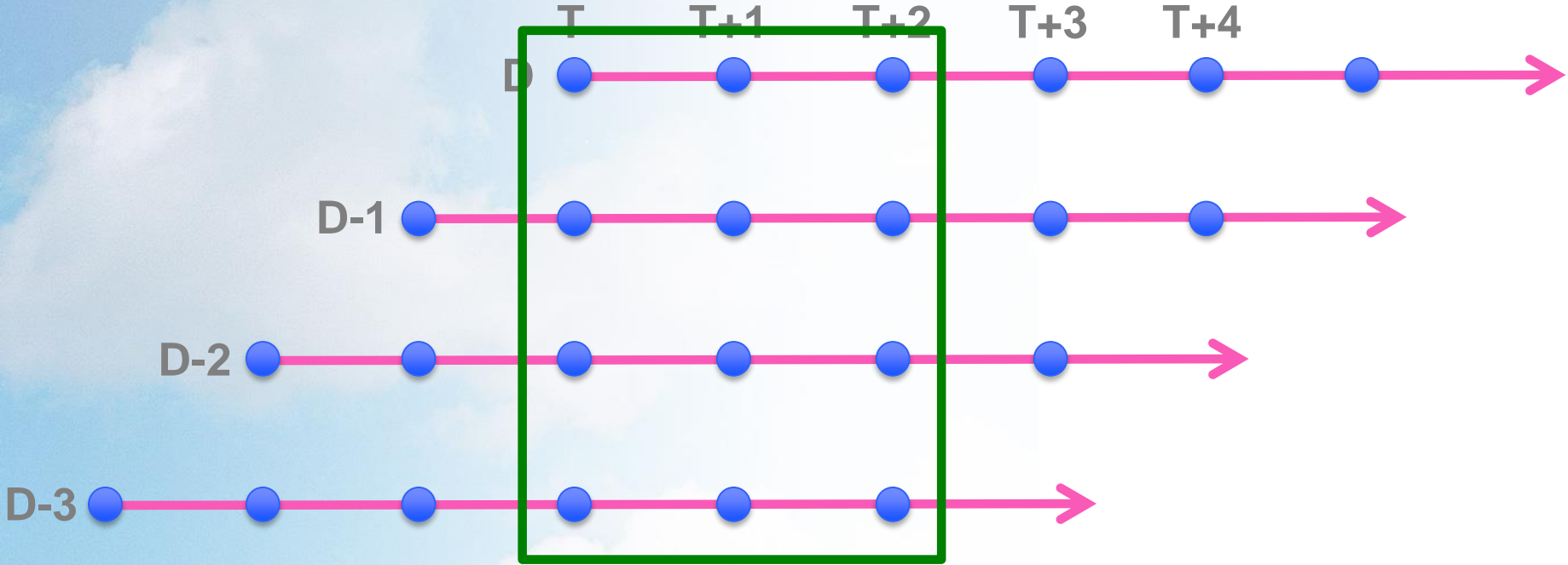
# Date & time



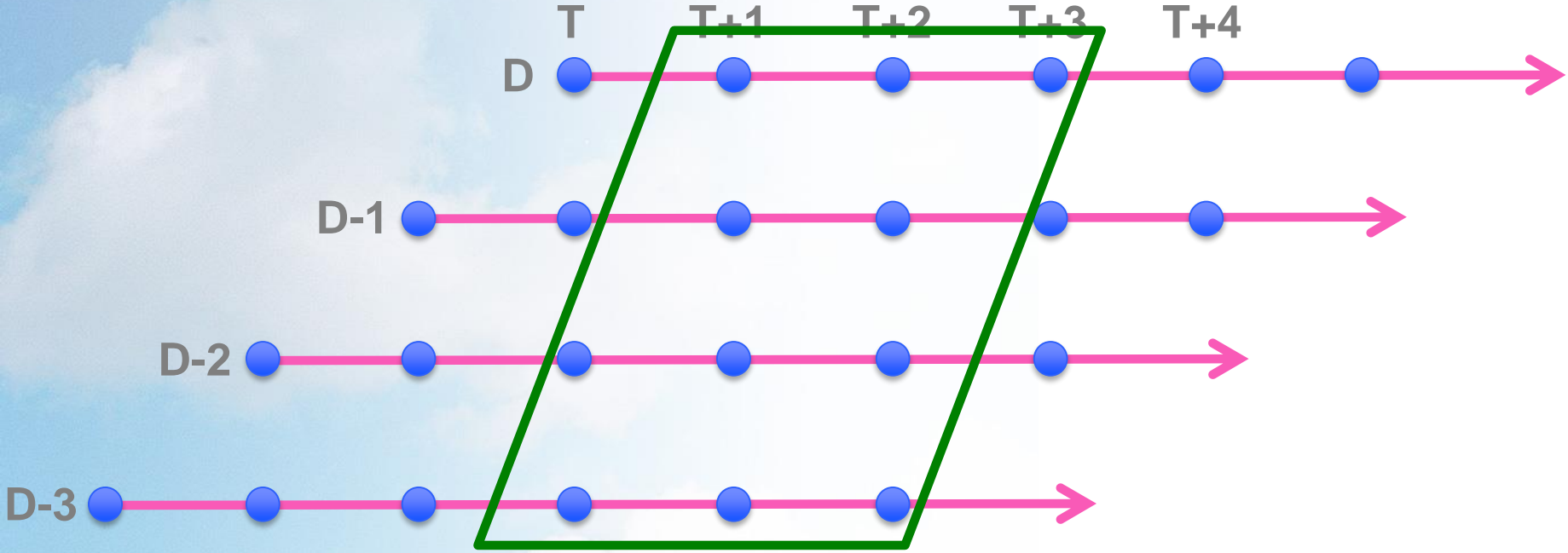
# Date & time



# Date & time

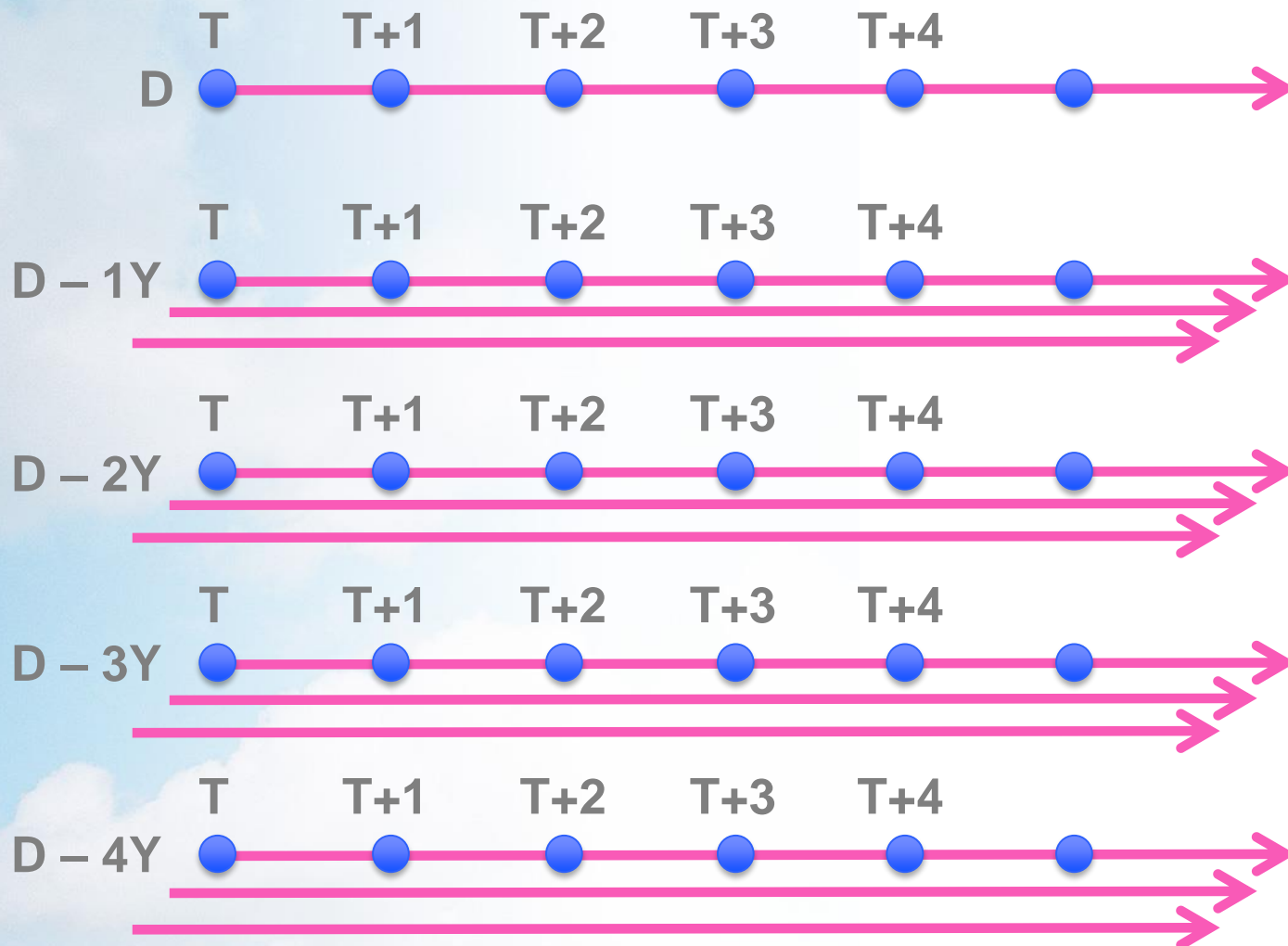


# Date & time

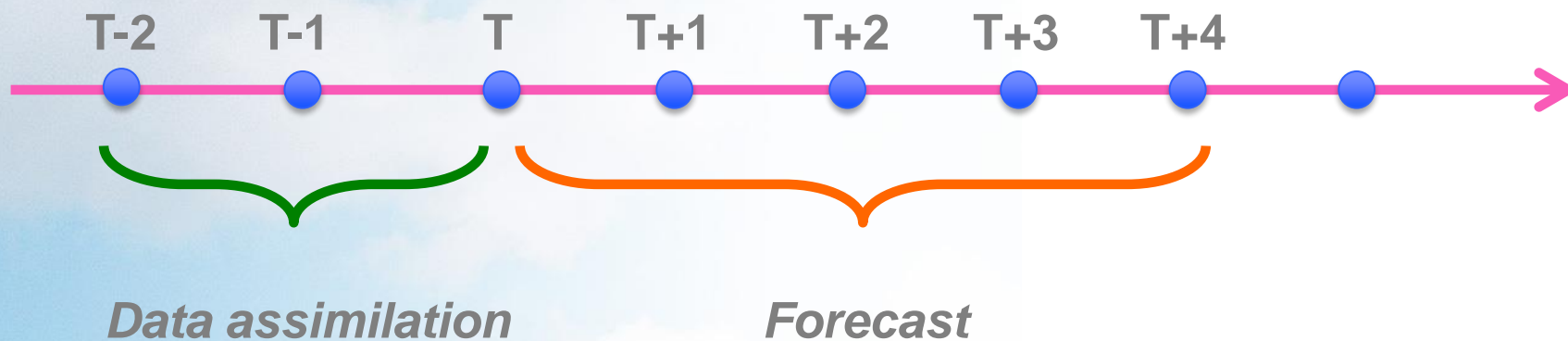




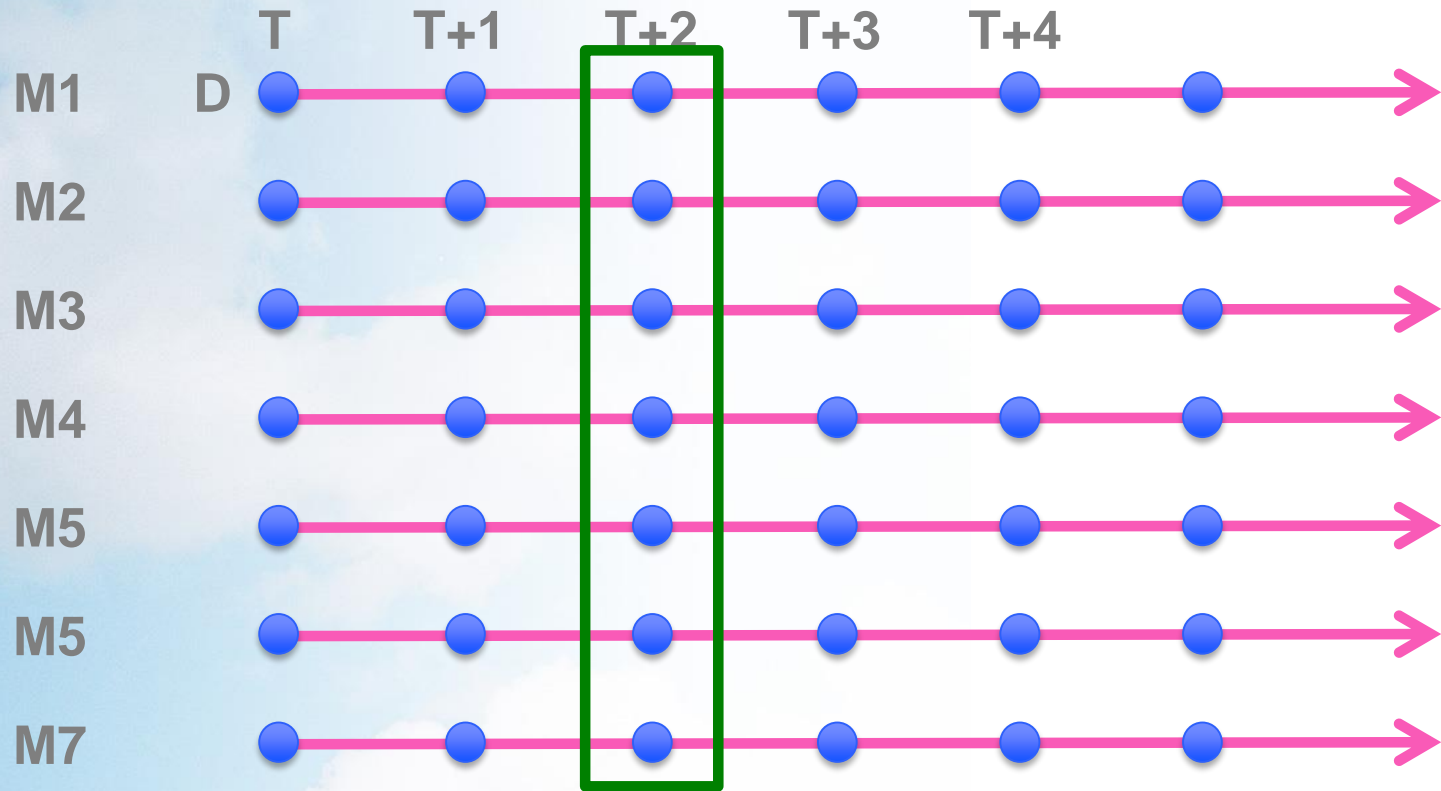
# Date & time (Hindcasts)



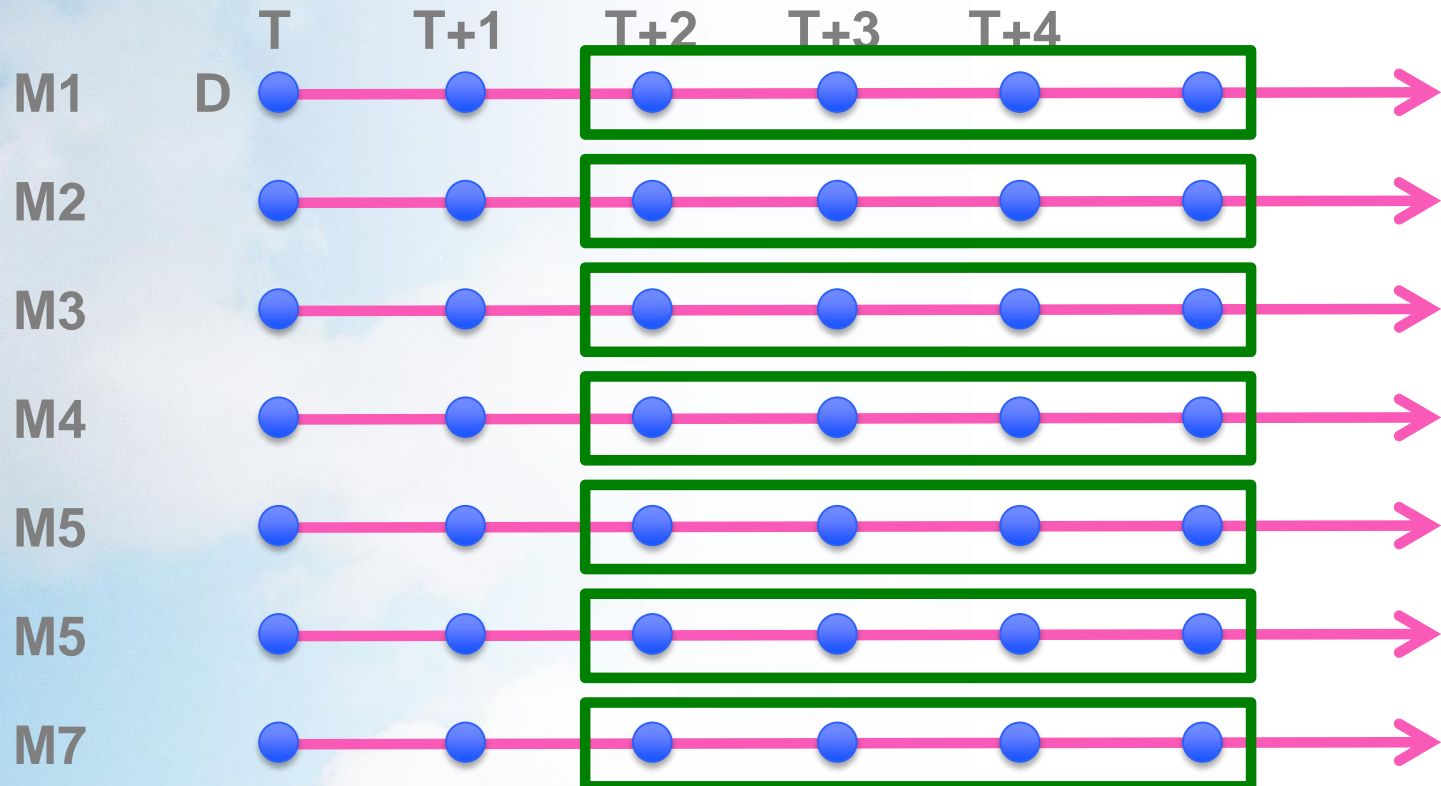
# Date & time (long window 4D-var)



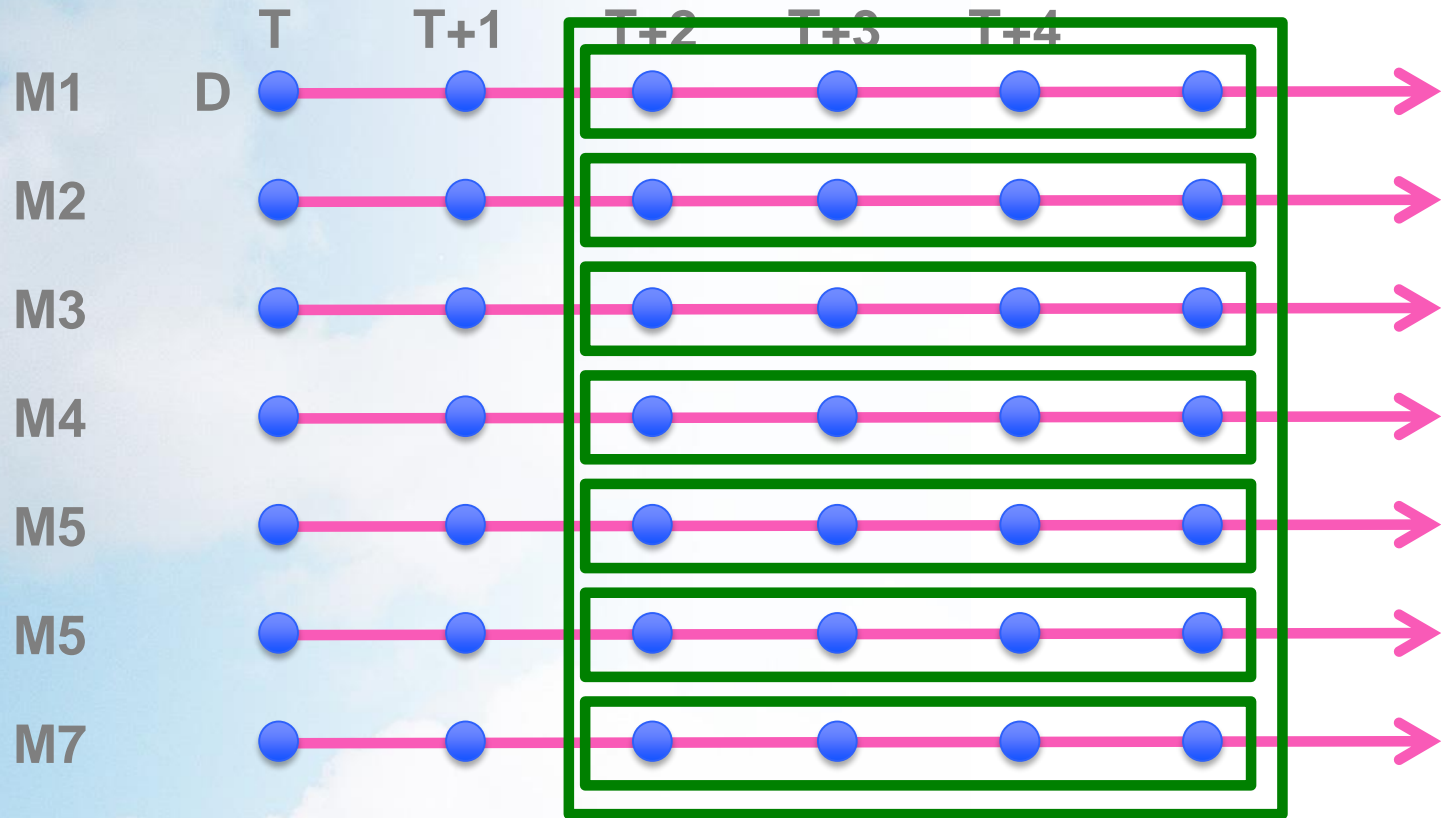
# Averaging: ensemble means



# Averaging: monthly means



# Averaging: monthly means of ensemble means



# Compression

# GRIB simple packing

---

- Maps floating point range  $[Field_{min}, Field_{max}]$  to integer range  $[0, 2^n - 1]$
- It's equivalent to sampling the field into  $2^n$  buckets
  - Packing is lossy
- $n$  can be anything between 1 and 32 (standard does not prevent  $n$  to be 255!)
- Most of the fields are packed with  $n = 16$ .
  - GRIB fields are half the size of the equivalent single precision float (or a quarter of double precision)
- Blind conversion from GRIB to NetCDF will create files twice as large (NC\_FLOAT) or four time bigger (NC\_DOUBLE)

# NetCDF supports “simple packing”

---

- Using **scale\_factor** and **add\_offset**, and packing to NC\_BYTE, NC\_SHORT, NC\_INT
  - Please note that these are signed (unsigned version comes with NetCDF4)
- Only multiple of 8 bits are supported
  - NC\_BYTE = 8, NC\_SHORT = 16, NC\_INT = 32
- Missing values:
  - GRIB uses a “bitmap” to mark the missing values
  - NetCDF uses **\_FillValue** to mark missing values
  - Consequence:
    - When packing NetCDF to NC\_BYTE or NC\_SHORT , we have 1 less value than GRIB
    - We cannot have encode the same range



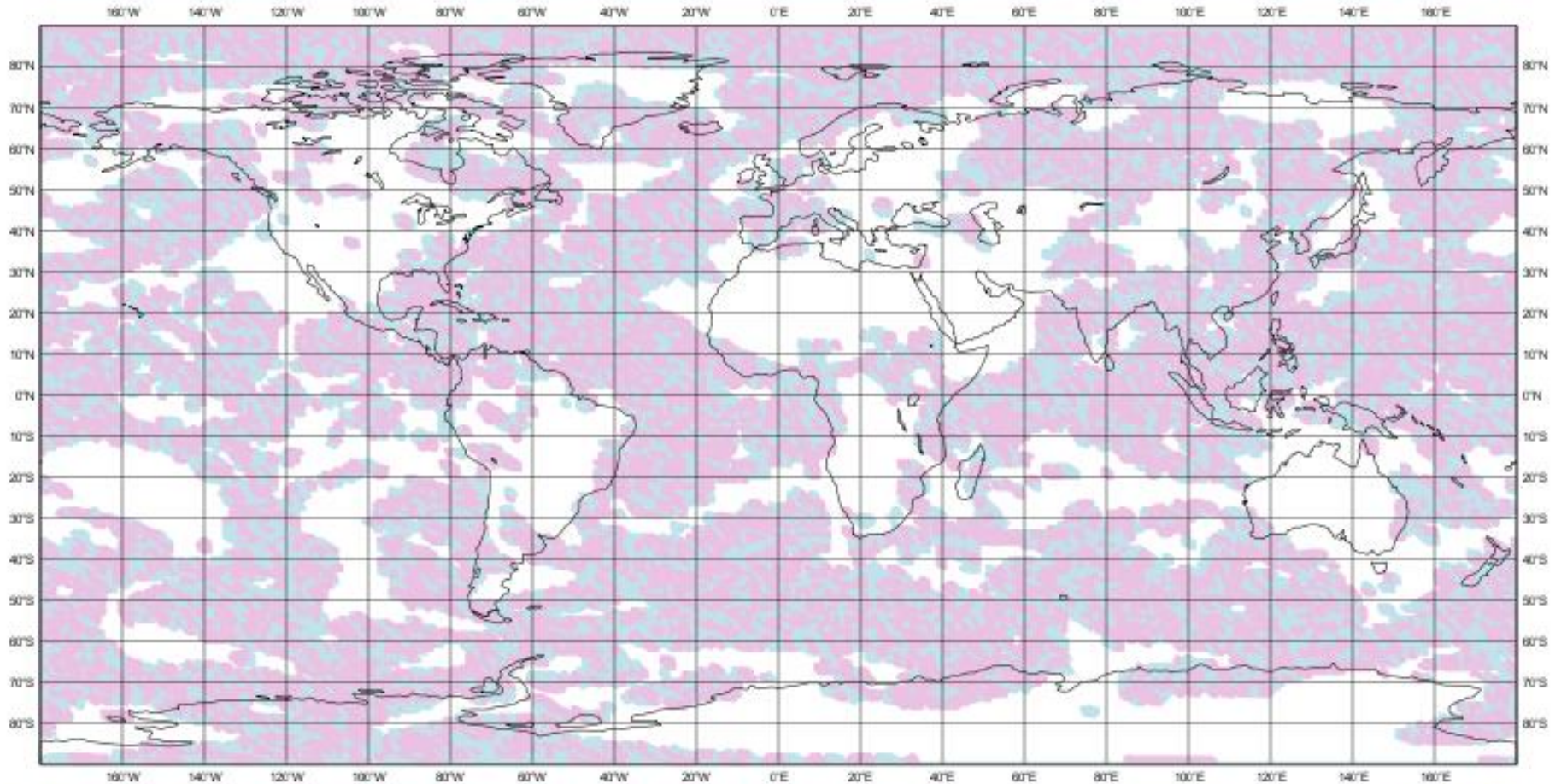
# Conversion and “simple packing”: major issue

---

- GRIB applies simple packing per 2D field
- NetCDF may apply packing per 3D (space and level, space and time) and even 4D fields (space, level and time)
- Consequence: mapping floating point range  $[\text{Field}_{\min}, \text{Field}_{\max}]$  to integer range  $[0, 2^n - 1]$  is done on more values in the case of NetCDF
  - Higher loss of accuracy
- Conversion leads to loss of information !!!!!
  - That’s not good™

# Difference is small, but non-zero (showing precipitations)

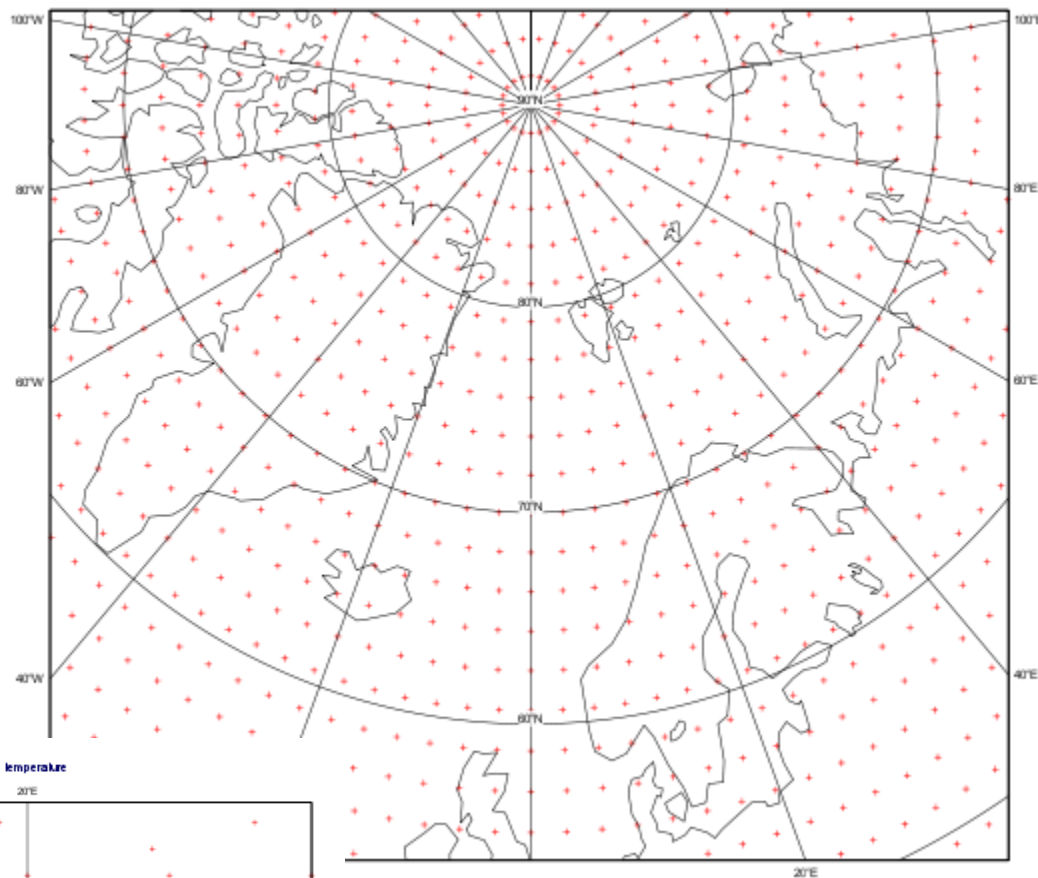
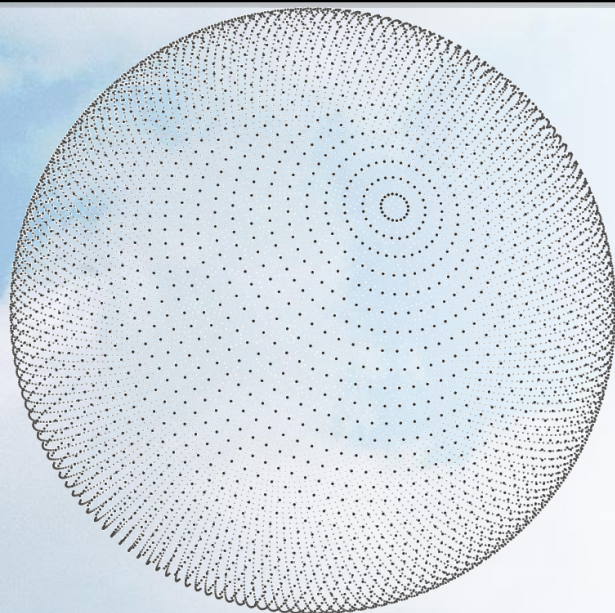
Difference GRIB NetCDF after conversion (MAX=5.60289e-07)...



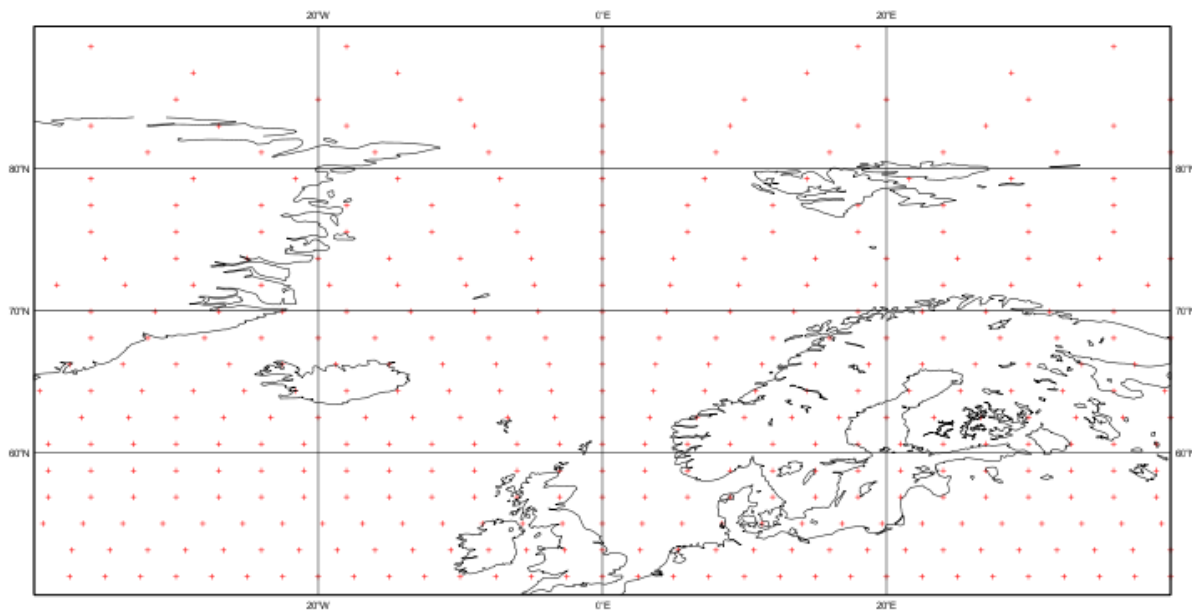
filesize: 2.22.7 (84 bit) - hydrat - mat - Tue Sep 23 17:14:18 2014

# Grids

# Reduced Gaussian Grid



Sunday 01 January 1995 12 UTC ECMWF Forecast 1-h0 VT Sunday 01 January 1995 12 UTC surface 2 metre temperature



# What do I want from this workshop?

---

- A general agreement on how to map GRIB to NetCDF (parameters, units, metadata, file structures,...)
  - So no one complains that we “are not doing it right”
- A general agreement on how we deal with future requirements (new grids, new parameters, ...)
  - Maybe a tighter collaboration between WMO and the CF community, like we did for OGC?