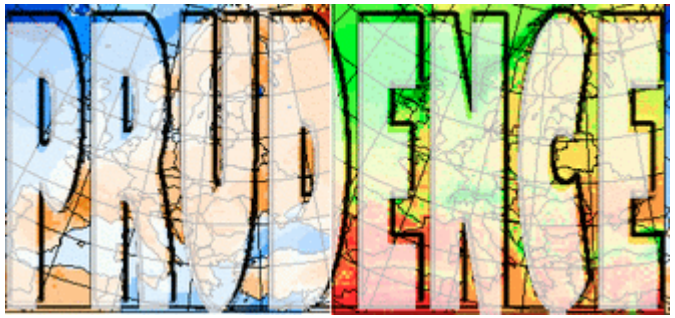
PRUDENCE and ENSEMBLES RCM archive experience

http://prudence.dmi.dk
http://ensemblesrt3.dmi.dk

Ole Bøssing Christensen
DMI

The PRUDENCE RCM database

- Data were created in the PRUDENCE project (2001-2004) with similar integration areas and a common specification of output variables
- Data received via DVD, tape, download, disks
- One canonical driving simulation supplemented by voluntary data sets with other drivers
- 21 surface fields available
- Around 500 GB data organized in 30-year files with daily data, plus monthly-mean files in native as well as 0.5 deg. regular grids
- Accessible through direct download or OpenDaP/DODS
- Large capacity but limited bandwidth
- Has been open to the public since 2004



The ENSEMBLES RCM database

- Data are created in the ENSEMBLES project (2004-2009) with common integration area and a common specification of output variables
- Data received mostly on USB disks, also download e.g. from ECMWF
- Wish-list of around 130 fields. Not all fields on the output list are present for all simulations
- Currently around 12 TB data organized in decadal files with daily or sub-daily data, plus monthly-mean files in native as well as 0.25 deg. regular grids
- Accessible through direct download or OpenDaP/DODS
- Large capacity but limited bandwidth
- Has been opened to the public this year, following the definition of an ENSEMBLES data policy
- Email addresses are asked for and saved, but no real security involved

DMI

ENSEMBLES

The ENSEMBLES RCM database

- ERA-40 based simulations in the archive:
  - 15 simulations in 50km resolution, 17 in 25km resolution
- Several transient climate change calculations have arrived
  - 1 simulation in 50km, 15 in 25km plus 3 sets of GCM output data
- Left to the individual groups to provide metadata about their model simulations
- We have continually underestimated the required size of the database. Voluntary contributions have been received, and RCM groups deliver increasingly exhaustive data sets. Shortly, we will enlarge the capacity from 13TB to 22TB
- In the end, around 3500 model years in 25km res.

DMI

The ENSEMBLES RCM database

- THIS IS LOW BUDGET
  - Hardware and current salary expenses carried by ENSEMBLES
  - In the future: Probably the DMI
- Hardware: Dual-CPU HP Proliant server
- Software: debian/apache. OpenDaP, netCDF, webalizer, NCO
- 16 disks, originally 500GB, now 1TB in a redundant RAID array. Currently 12 more disks on order. Total hardware costs around 25 k€
- Piggyback on the DMI 100MBit/s network connection, and backup in the DMI mass storage system (IBM)
- Data massage, file manipulation etc. requires 0.2-0.5 persons depending on ”season”. Mostly shell scripting like renaming of files, and NCO command line manipulation of attributes. Moving around stuff between almost-full partitions
- Hence: User- and supplier-based debugging

DMI

The progress from PRUDENCE to ENSEMBLES

- A more varied GCM-RCM matrix
- More output fields
- Higher spatial and temporal resolution
- Transient runs and not time slices
- Also ERA40-based simulations
- Higher degree of standardisation

PRUDENCE GCM-RCM Matrix

	CNRM	DMI	ETH	GKSS	HC	ICTP	KNMI	MPI	SMHI	UCM
A2+HadAM3H		3	1	1	3	1	1	1	1	1
A2+ECHAM4		1							1	
A2+ARPEGE3	1									
B2+HadAM3H					1	1			1	1
B2+ECHAM4		1							1	
B2+ARPEGE3	3									

ENSEMBLES GCM-RCM Matrix

Global model Regional model	METO-HC Normal sens.	METO-HC Low sens.	METO-HC High sens.	MPIMET	IPSL	CNRM	NERSC	CGCM3	Total number
METO-HC	**1950-2100**	**1950-2100**	**1950-2100**	1950-2100 (08/2009)					4
MPIMET				**1950-2100**	1950-2050* (04/2009)				2
CNRM						**1950-2050**			1
DMI				1950-2050* (04/2009)		**1950-2100**	1950-2100* (06/2009)		3
ETH	**1950-2100**								1
KNMI				**1950-2100**					1
ICTP				1950-2100 (03/2009)					1
SMHI		1950-2100* (03/2009)		**1950-2100* (50km)** 1950-2100* (25km) 03/2009			**1950-2100**		3
UCLM	**1950-2050**								1
C4I			**1950-2100***	**1950-2050*** (A2)					2
GKSS**					1950-2050* (??)				1
Met.No**							**1950-2050***		1
CHMI**						1950-2050* (??)			1
OURANOS**								**1950-2050***	1
Total (1950-2050)	3	2	2	7	2	3	3	1	23

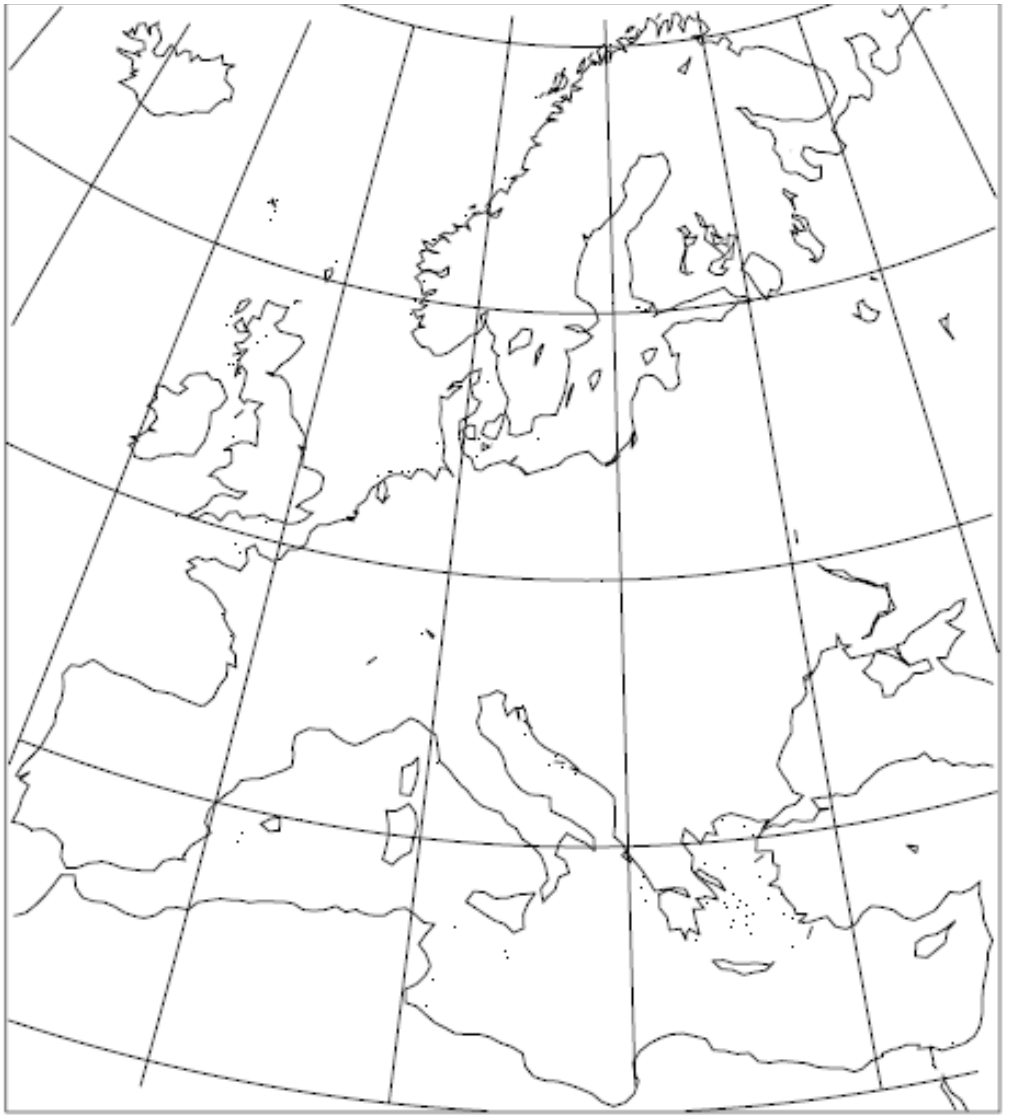
DMI

Currently 1900 model years, eventually 3100 years
-additional contributions expected…

*: non-contractual runs **:affiliated partners without obligations

Common integration area

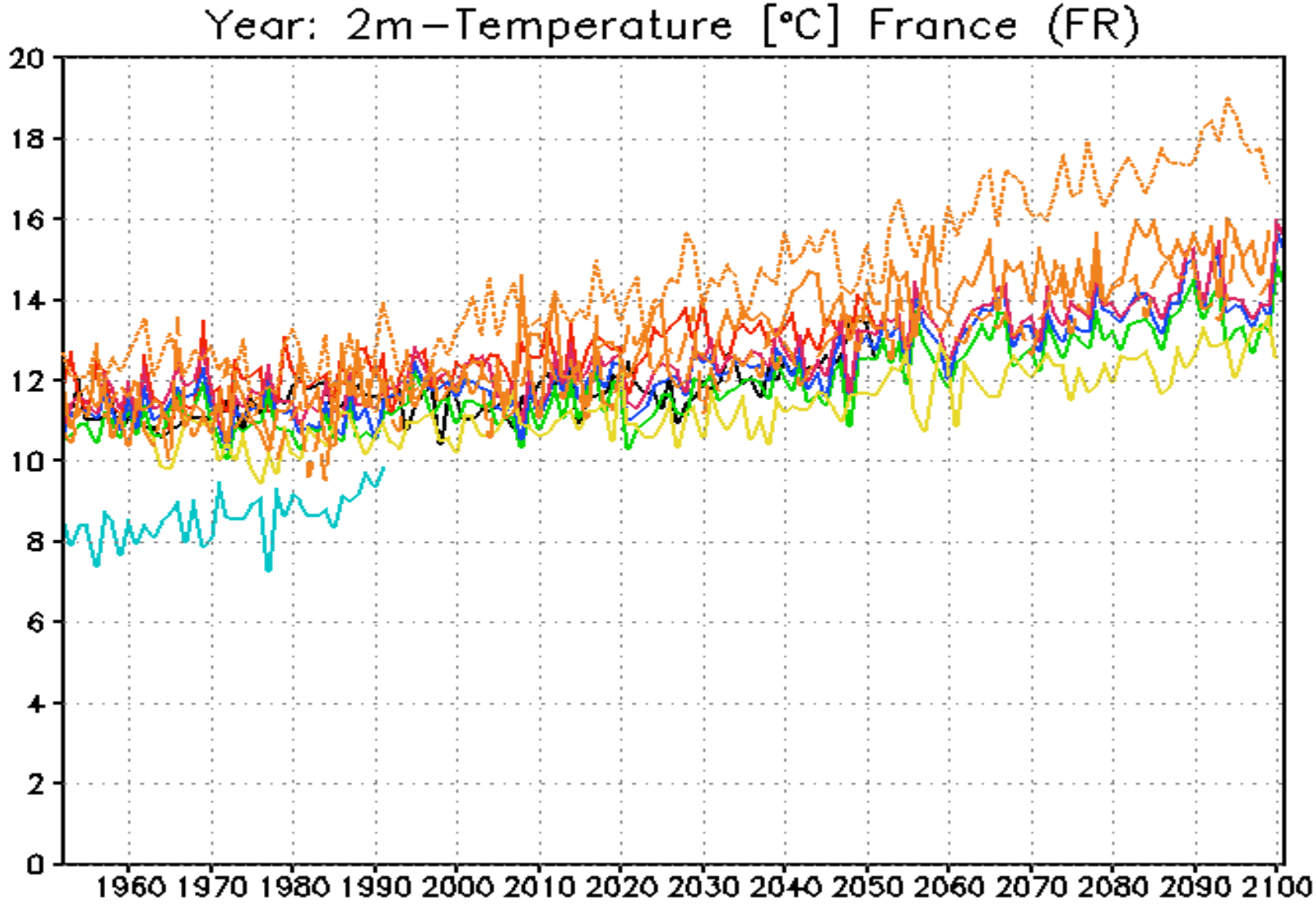
- Internal zone
- Models with rotated lat/lon
- Common area:
- [lon1,lon2,lat1,lat2,polon,polat]=[-21.72,15.46,-20.668,20.90,18.00,-39.25]
- nlon x nlat = 170 x 190



Common output field list

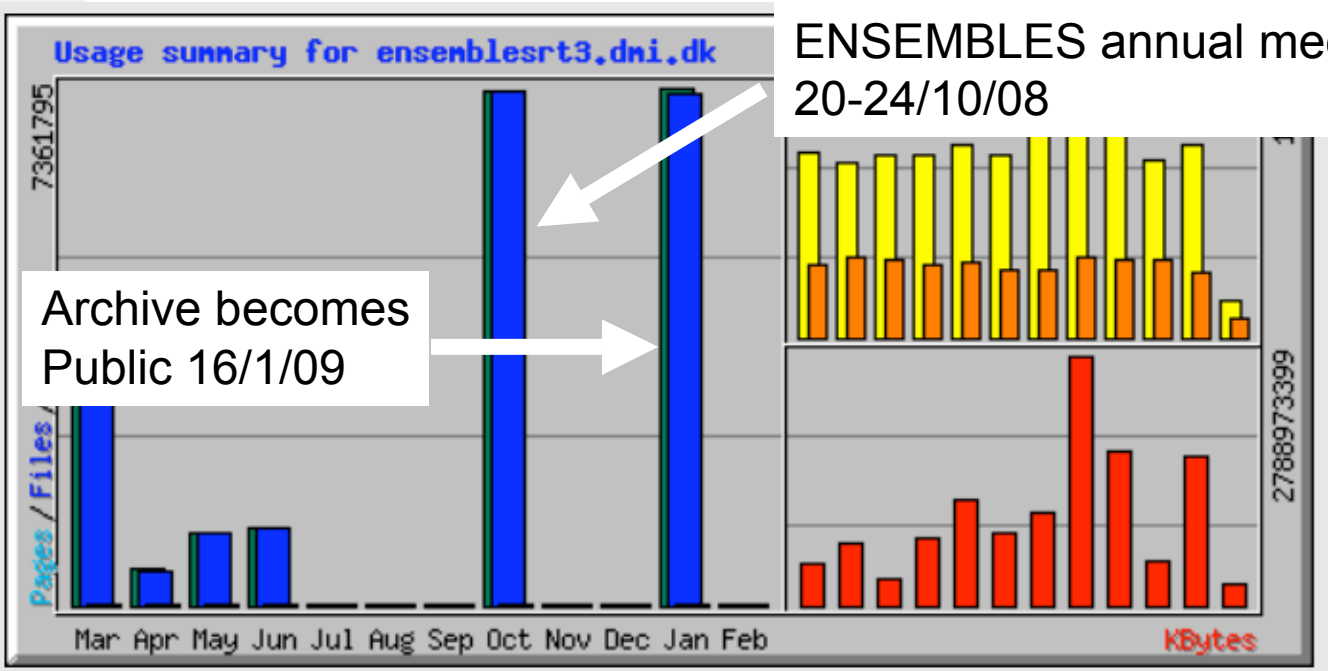
12	**Interpolation to CRU-like 0.25-degree grid: All 3d-fields are interpolated bilinearly without discrimination between land or sea. All surface fields are interpolated land points to land point and sea points to sea point**								quantity is not available in your simulation.
13				RT3	Optional				
14		short name		(to be submitted to the data	(to be stored				
15		(PCMDI	MKS Units	server)	locally)			CF	Provider's
16	Field name	+ extensions)	CF Standard			Remarks	Cell_methods	standard_name	Remarks
17									
18	**1) Near-surface data**								
19									
20	2-meter temperature	tas	K	dci	1h/3h		time: mean	air_temperature	
21	Daily maximum 2-m temperature	tasmax	K	dc			time: maximum	air_temperature	
22	Daily minimum 2-m temperature	tasmin	K	dc			time: minimum	air_temperature	
23	Max Surface temperature (soil)	tsosmax	K	dc	1h/3h		time: maximum	soil_temperature	
24	Min Surface temperature (soil)	tsosmin	K	dc			time: minimum	soil_temperature	
25	Surface temperature	ts	K	dci	1h/3h		time:mean	surface_temperature	
26	Sea surface temperature	sst	K	d at 0 UTC				sea_surface_temperature	
27	10-meter U wind	uas	m s-1	dci	1h/3h		time: mean	eastward_wind 1)	
28	10-meter V wind	vas	m s-1	dci	1h/3h		time: mean	northward_wind 1)	
29	10-meter wind speed	wss	m s-1	dci	1h/3h		time: mean	wind_speed	
30	10-meter daily max. wind speed, without gust	wssmax	m s-1	dc			time: maximum	wind_speed	
31	10-meter daily max. wind speed incl. gust	wsgsmax	m s-1	dc			time: maximum	wind_speed_of_gust	
32	2-meter specific humidity	huss	kg kg-1	dci	1h/3h		time: mean	specific_humidity	
33	2-meter relative humidity	hurs	1	dci	1h/3h		time: mean	relative_humidity	
34	Daily maximum 2-m relative humidity	hursmax	1	dc			time: maximum	relative_humidity	
35	Daily minimum 2-m relative humidity	hursmin	1	dc			time: minimum	relative_humidity	
36	2-meter dew point temperature	tdps	K	dci	1h/3h		time: mean	dew_point_temperature	
37	Mean sea level pressure	psl	Pa	dci	1h/3h		time: mean	air_pressure_at_sea_level	
38	Surface pressure	ps	Pa	dci	1h/3h		time:mean	surface_air_pressure	
39	Surface pressure	ps	Pa	6h				surface_air_pressure	

The ENSEMBLES Ensemble



Organisation

- Debian stable and apache 2
- Gzip'ed netCDF files following the CF1.0 convention for
  - Direct download
  - OpenDaP access
    - Sub-windows in space and sub-periods in time
    - Strides
    - Direct integration in "OpenDaP-aware" applications
- Daily, sub-daily and monthly-mean files on native grid
- Common regular interpolated grid for monthly means as well
- Not very user friendly! No online previews or analyses (yet)



Summary by Month

Month	Daily Avg				Monthly Totals					
	Hits	Files	Pages	Visits	Sites	KBytes	Visits	Pages	Files	Hits
Feb 2009	1350	919	395	48	98	234332345	194	1582	3676	5400
Jan 2009	237477	233992	245	33	345	1668234809	1045	7625	7253775	7361795
Dec 2008	579	398	228	30	413	508824151	954	7086	12346	17956
Nov 2008	1128	837	431	44	423	1728716595	1349	12932	25127	33856
Oct 2008	236179	235234	364	41	435	2788973399	1274	11310	7292263	7321574
Sep 2008	1018	494	460	36	359	1037260176	1107	13824	14841	30549
Aug 2008	768	636	206	31	371	823956173	982	6389	19724	23835
Jul 2008	602	462	228	33	402	1190587004	1033	7083	14337	18681
Jun 2008	36796	36652	276	32	386	749414979	988	8284	1099580	1103890

Usage

Polish Res. Ctr. Agricult. Env.
bvbs-wan.de
ZMAW
AUTH
UK Met Office
Météo-France

DMI

Top 10 of 345 Total Sites By KBytes

#	Hits		Files		KBytes		Visits		Hostname
	4604	0,06%	4544	0,06%	662290958	39,70%	6	0,57%	150.254.171.10
	2233	0,03%	2192	0,03%	307004312	18,40%	12	1,15%	195.37.166.251
	9435	1,35%	1119	0,02%	303069757	18,17%	0	0,00%	136.172.85.128
	552	0,01%	533	0,01%	66644326	3,99%	6	0,57%	155.207.124.26
	902	0,01%	450	0,01%	52633352	3,16%	24	2,30%	151.170.240.10
	456	0,01%	456	0,01%	40519390	2,43%	0	0,00%	150.254.171.51
	227	0,00%	182	0,00%	37730693	2,26%	17	1,63%	137.129.13.91
8	379	0,01%	209	0,00%	37123223	2,23%	11	1,05%	137.129.13.90
9	638	0,01%	540	0,01%	35028841	2,10%	38	3,64%	145.23.254.155
10	231	0,00%	167	0,00%	34591458	2,07%	6	0,57%	161.67.122.236

Usage

- Only logging for a month
- Mostly project partners
- -But also other institutions

Now what?

- Several 2nd-tier simulations and West African simulation data on their way
- Large interest in data availability
- The existing setup works fine as a major data centre, provided we have
  - No order-of-magnitude increase in traffic (currently we use a few % of the theoretial maximum DMI non-operational bandwidth)
  - No large future additions
  - Users like modellers, who can do without online previews and exhaustive documentation

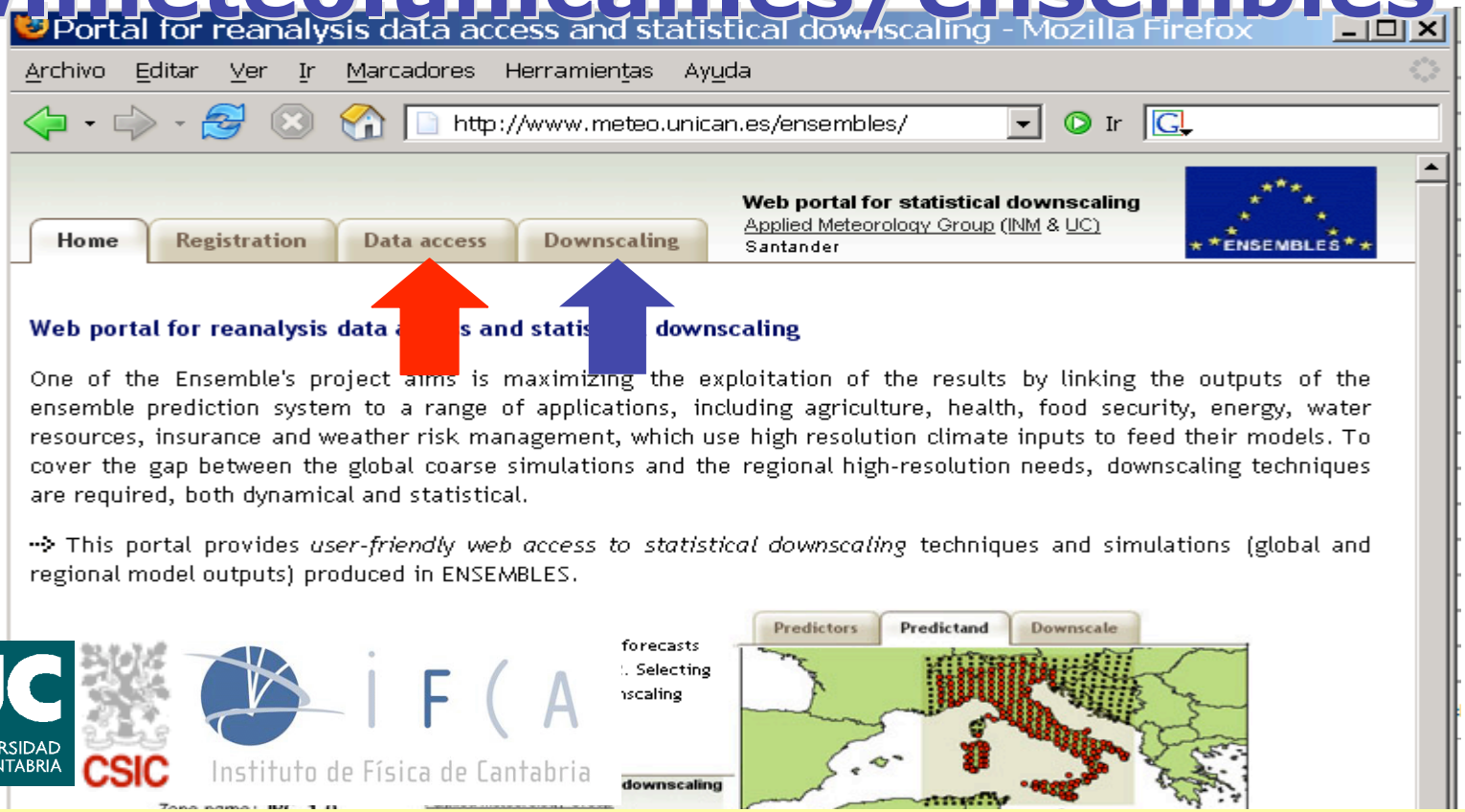
Needs for the next project

- Let's guess:
  - 10 areas
  - All 50km resolution
  - 150year runs
  - 50 runs per area
  - Same fields as in ENSEMBLES
- For us, 1.900 model years take up 12TB
- So, a complete data center needs 10*150*50/4 * 12TB/1900 = 120TB if constructed like ENSEMBLES
- -An order of magnitude more than ENSEMBLES, but still, this is not hard and is relatively cheap
- Mind you: more than 100 fields per expt.!

The ENSEMBLES sd portal

Slides from José Manuel Gutierrez

The ENSEMBLES sd portal: www.meteo.unican.es/ensembles



Data Access Portal: Web portal to access a reduced set of predictors and observations (locally stored).

Statistical Downscaling Portal: Web portal to run statistical downscaling on gridded or local observations.

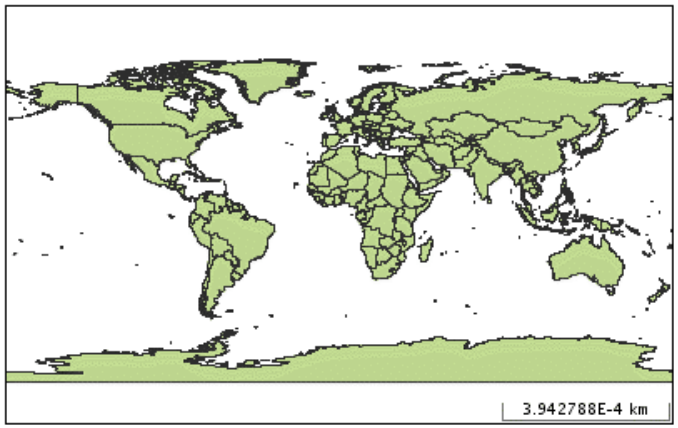
GCM Data Availability. Seasonal & ACC

• **DEMETER**. Multi-model seasonal prediction experiment including seven models ran for six months four times a year using 9 different perturbed initial conditions (9 members).

• **ENSEMBLES** Stream 2. Just delivered, establishing a connection via OpenDAP (in this case the data is accessed remotely).

Daily worldwide datasets obtained from different sources:

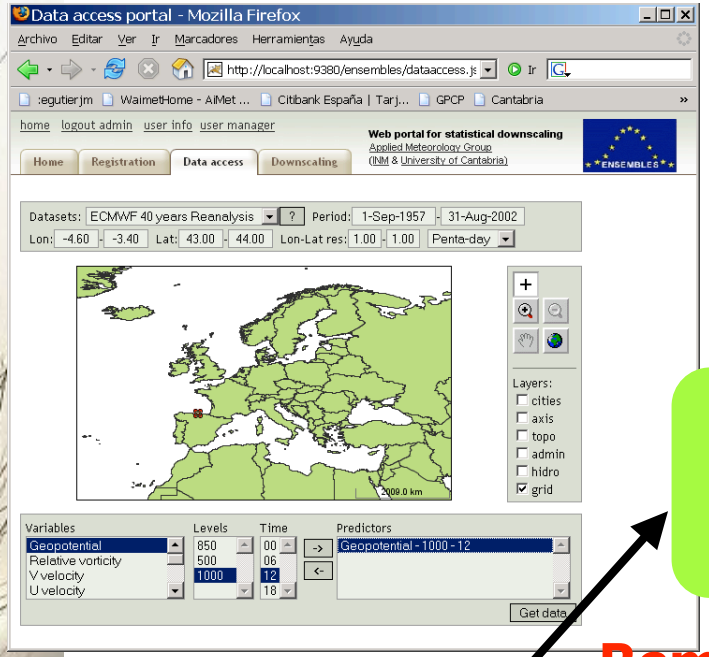
- **CERA (ENSEMBLES repository)**
- **PCMDI**
- **Local Providers.**



Some examples:

- **CERA_MPI-ECHAM5,** including **20th century** data (1961-2000) and scenarios **A1B**, **B1**, and **A2** (2001-2100).
- **CNRM-CM3 (local provider),** including **20th century** (1961-2000) and scenarios **A1B**, **B1**, and **A2** (2001-2100).

Distributed Data Access.
GRID Computing



Distributed-data Application

Downs. Portal

OpenDAP Client

Remote OpenDAP **http**

Local access to data

OpenDAP Servers

Data (local)

Data (ECMWF)

Data (DMI)

Data (remote)

Enabling Grids for E-sciencE

GRID COMPUTING

We are participating in existing E-Science EU initiatives (EGEE Earth Science VO) in order to move forward to GRID computing to facilitate distributed resources sharing.

public.eu-egee.org

PCMDI is already involved in GRID.