The storage features and needs for numerical modeling at ARPA FVG - CRMA

Scientific data management approaches,
data analysis and tools

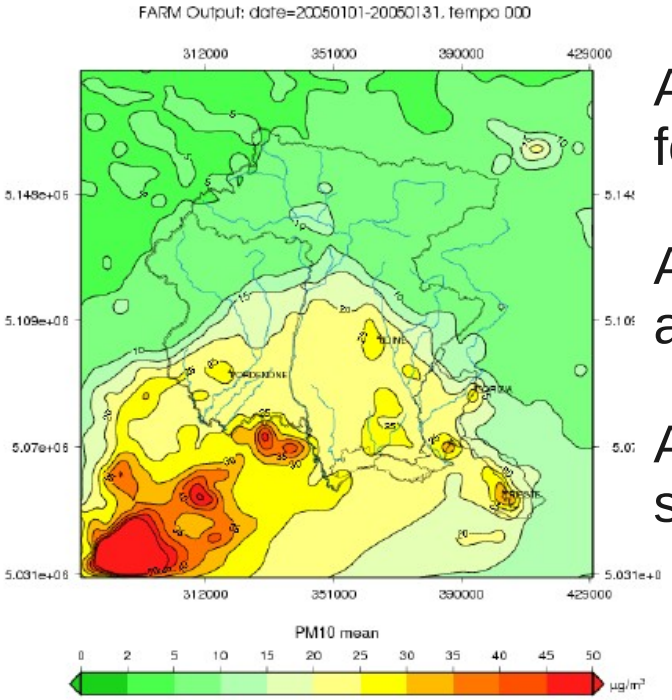
Trieste (ICTP)
September 05, 2013

ARPA FVG – CRMA
Centro Regionale di Modellistica Ambientale
crma@arpa.fvg.it

Outline of the presentation

- Main computational activities at ARPA FVG – CRMA
- Typical use of model outputs and updates frequency
- Amount of data generated yearly and needed disk space
- Criteria of data selection for permanent data storage
- Strategies adopted at CRMA for data storage

Main computational activities at CRMA

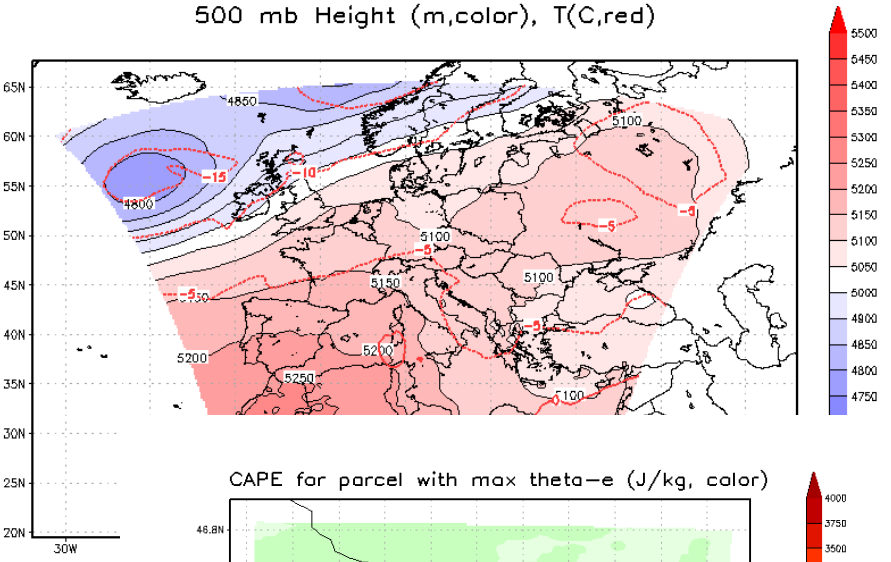


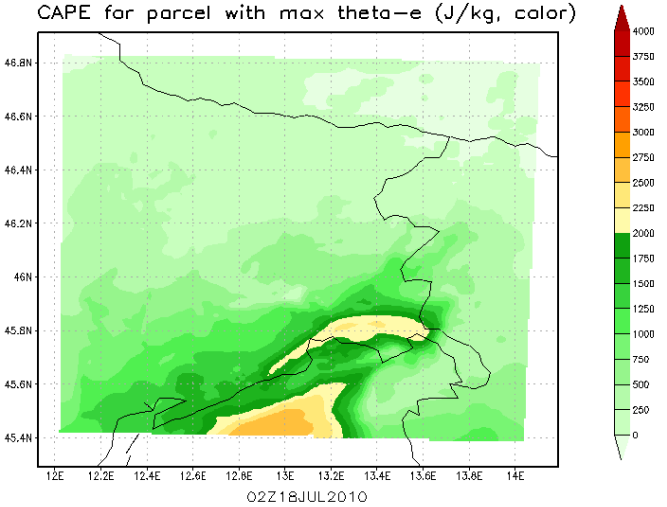
Air quality forecasts,

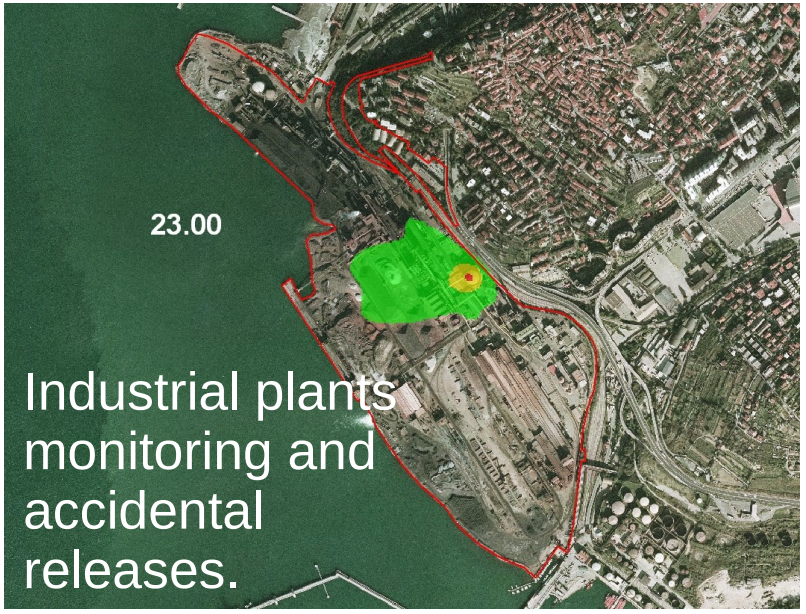
Air quality analysis,

Air quality scenarios.

High resolution weather forecasts and weather analysis.

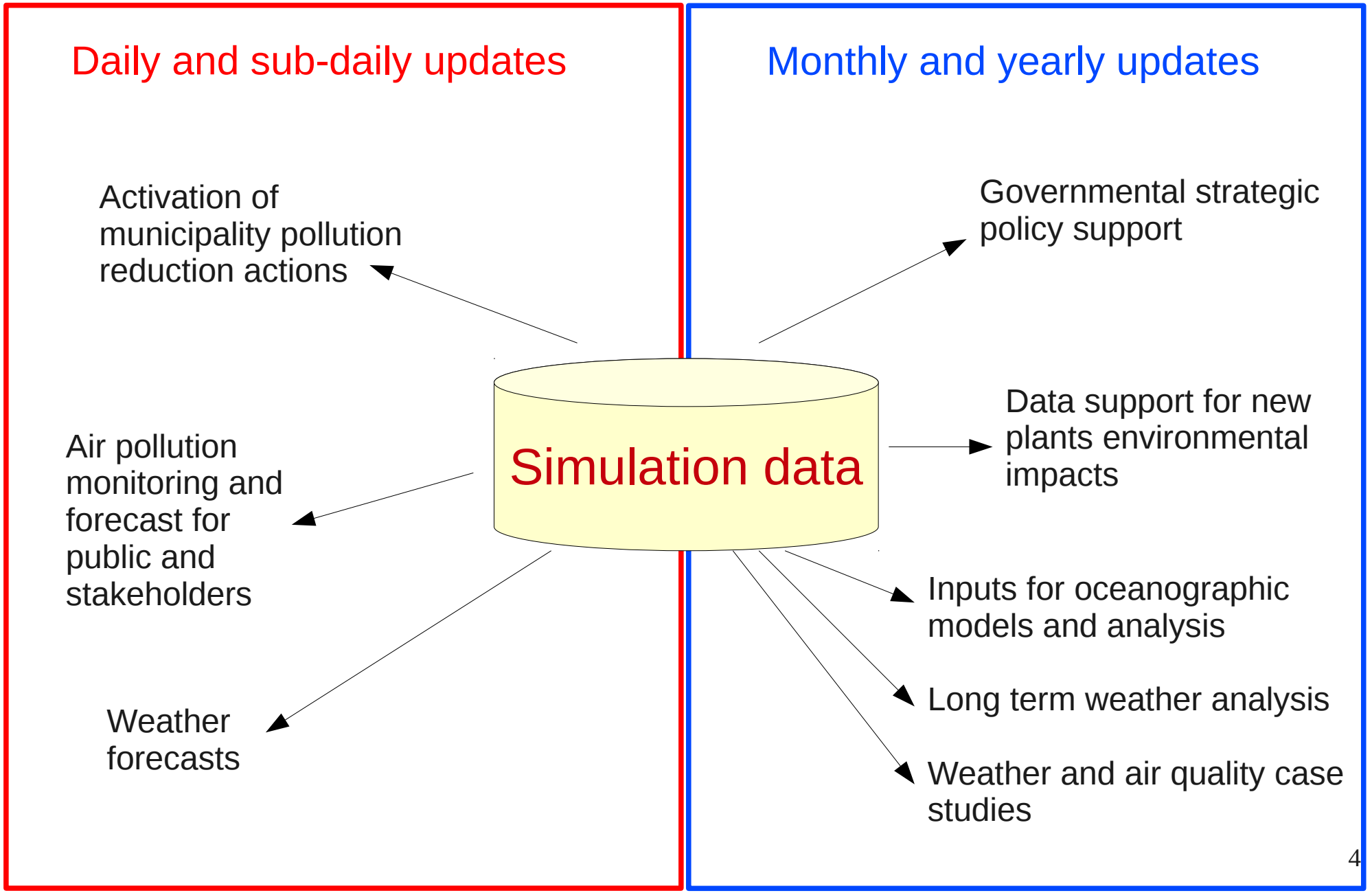






Industrial plants monitoring and accidental releases.

Typical use of model outputs ad update frequency



Amount of data generated for each update and sum over one year

Daily and sub-daily updates

Activation of municipality pollution reduction actions
0.5 GB/day

Air pollution monitoring and forecast for public and stakeholders
2.5 GB/day

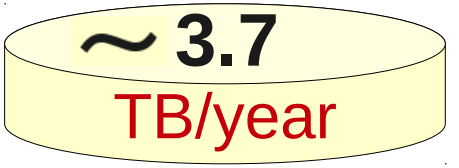
Weather forecasts
5.0 GB/day

Monthly and yearly updates

	updates
Governmental strategic policy support **180 GB/policy**	**1/year**
Data support for new plants environmental impacts **15 GB/plant**	**6/year**
Inputs for oceanographic models and analysis **12 GB/(oceanographic year)**	**5/year**
Long term weather analysis **500 GB/(meteorological year)**	**2/year**
Case studies: weather **25 GB/(case study)** air quality **20 GB/(case study)**	**10/year**

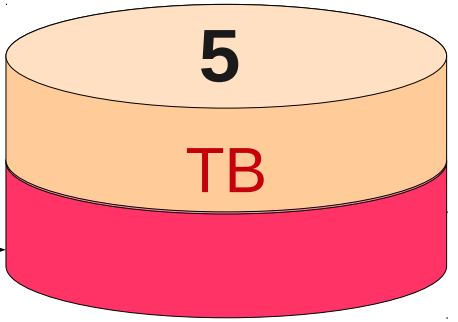
Yearly disk space required and limits of a foolish archival

Total amount of simulated data per year

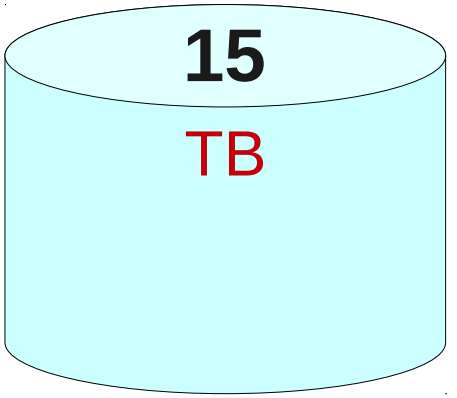


ARPA FVG – CRMA total availability now,

half already filled



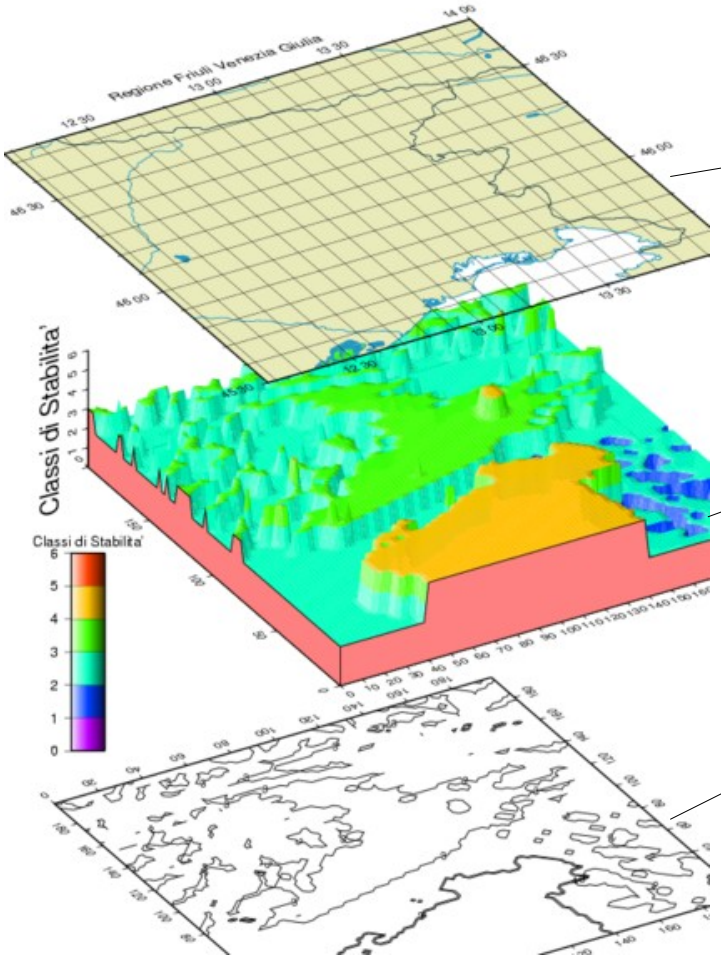
ARPA FVG – CRMA expected storage availability increase for the next 3 years

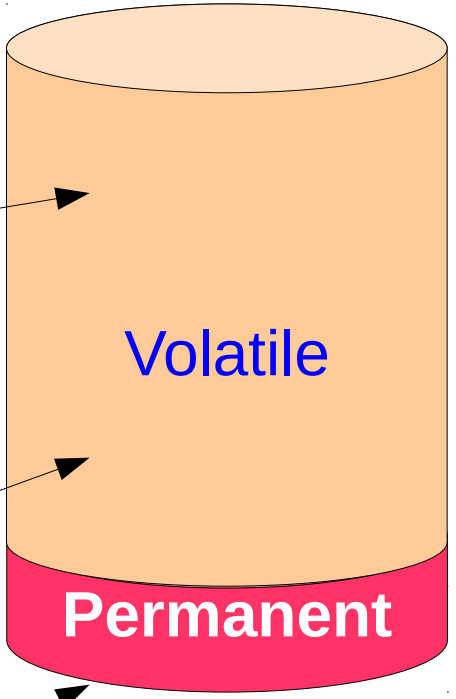


In archiving all simulated data, the system will collapse in less than one year

Relevance of simulated data and archival criteria

A numerical model requires all produced data during the simulation of the reality, but **a subset only** is used **for results interpretation and applications**.





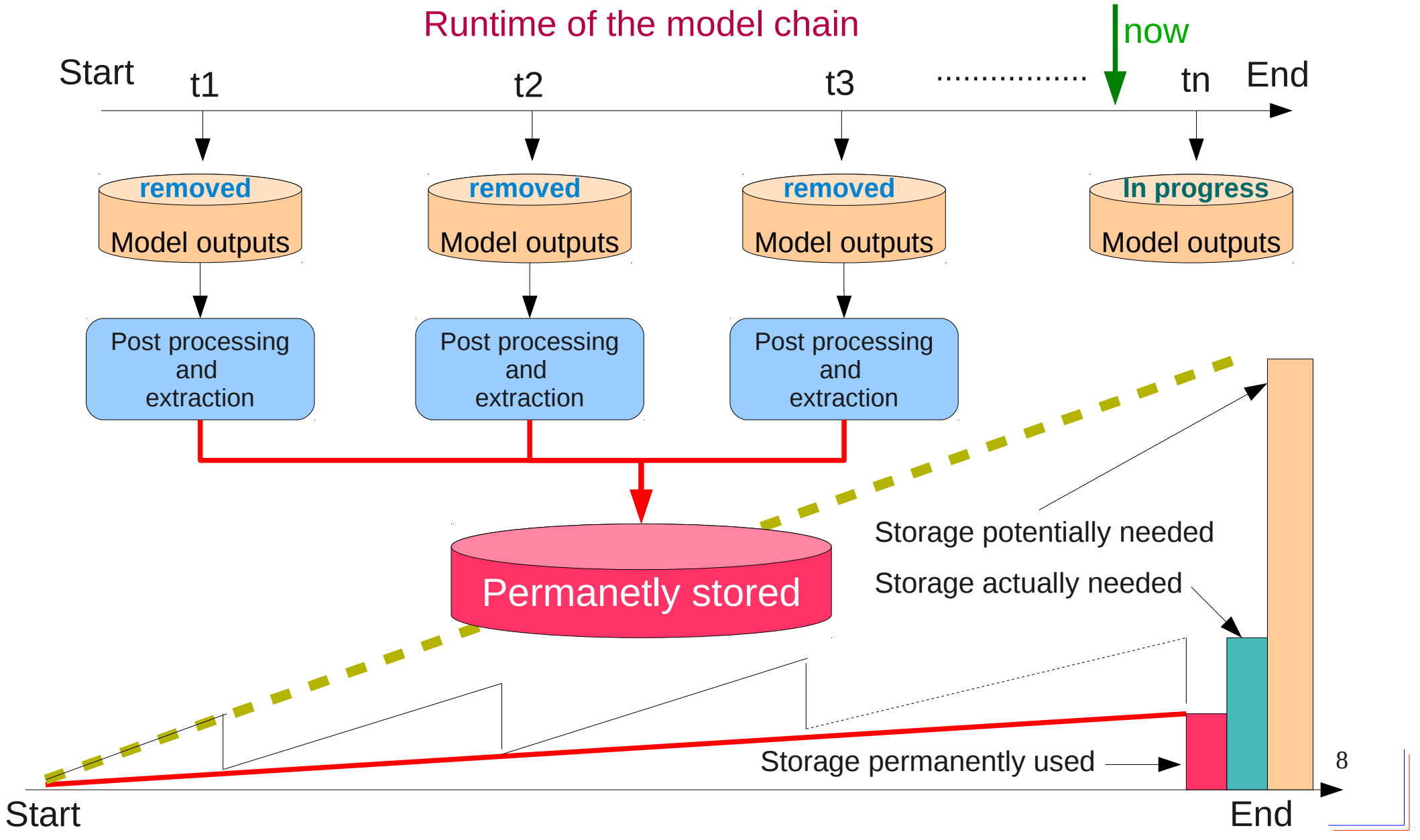
Fields selected as a function of:

- probability of future post processing
- permanent storage availability for the next 10 years

Only a small subset of all generated bytes are permanently stored, most of the fields are removed according a time window specifically defined for the volatile part

On the fly post processing and related data archival criteria

There are simulations that require huge disk space to be completed, but post processing can be performed during the run, so removal of not needed data is allowed before the run end



No simulation storage, instead rerun model

In many cases the simulations are **executed very fast**, i.e. less than one computational day, and the **outputs** are **requested** for postprocessing **once or less in a year** so it is convenient to **rerun** the model instead to fill the storage

Cornerstones of this approach are:

- Permanent storage of all the inputs necessary to run the model
  (i.e. boundary and initial conditions, sources of pollutants, meteorology, etc.)

In practice → Standardization of data formats required by models (i.e. netCDF, GRIB, etc)

- Collection and storage of metadata on the computational environment
  (i.e. hardware, compilers, libraries, models version, post processors versions, etc.)

In practice → Classification of each simulation (i.e. 0141F0B0B1_2005) and use of archival software support (i.e. WikiCRMA, may be SISCO in the next future)

- Definition and practical implementation of a procedure for the run preparation, execution and post processing (minimize time and human errors in preparation and data handling)
  (i.e. where to find the inputs, how to prepare the run parameters, which kind of post processed fields to produce etc.)

In practice → Workflow implementation (home made scripts at CRMA are currently in migration toward ecFlow and Kepler)

Permanent archival of data

The **permanent** archival of data may be in different **hardware devices** and **compression** according to:

Frequency of data retrieval + amount of archived data

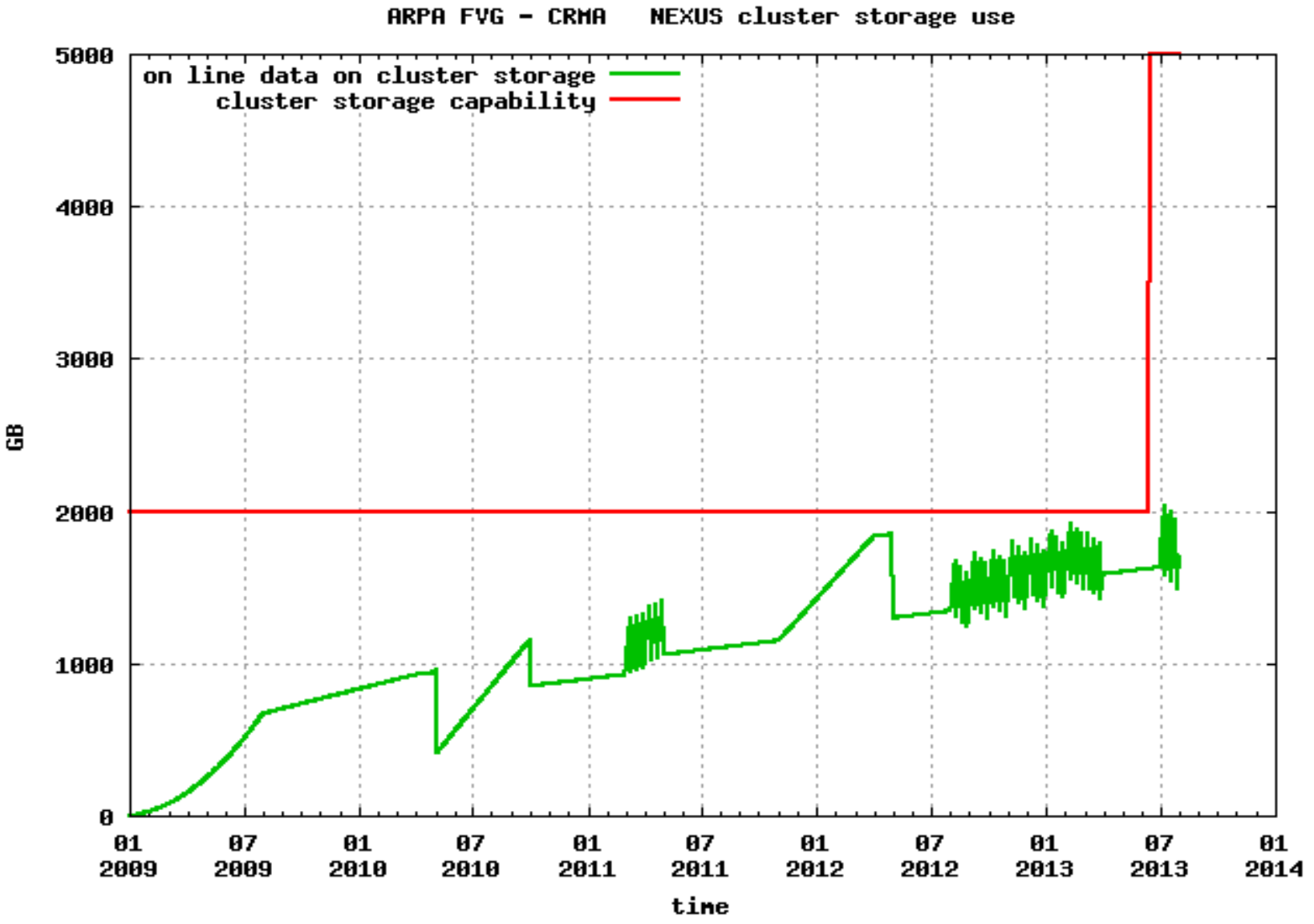
Simulation outputs:

Frequency of data retrieval	Data amount	Compression	Device
< 0.5/year	< 15 GB	Yes/Not	DVD
< 0.5/year	> 15 GB	Yes/Not	External disk
(1 or 2)/year	any	Yes	Cluster storage on line
> 2/year	any	Not	Cluster storage on line

Cluster NEXUS storage history

This is the end of the presentation

Thanks



Supplementary information

Example of data volumes generated by WRF model

durata	dominio 1	dominio 2	dominio 3	totale
1 giorno	127 MB	529 MB	569 MB	1.35 GB
11 giorni	1116 MB	5116 MB	5899 MB	12.52 GB
30 giorni	5256 MB	17567 MB	17739 MB	40.56 GB
366 giorni	64126 MB	214312 MB	216414 MB	494.85 GB

Tabella 3.3: Memoria occupata dai file di output del WRF.