Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Data quality assurance: general overview and the EC-Earth case

PA Bretonnière, EC-Earth meeting, 21/05/2019

Outline

1. Quality Assurance overview
   a. Why checking the data?
   b. Sources of errors
   c. What is a good dataset?

2. QA project and software inventory
   a. C3S512
   b. Existing data checkers

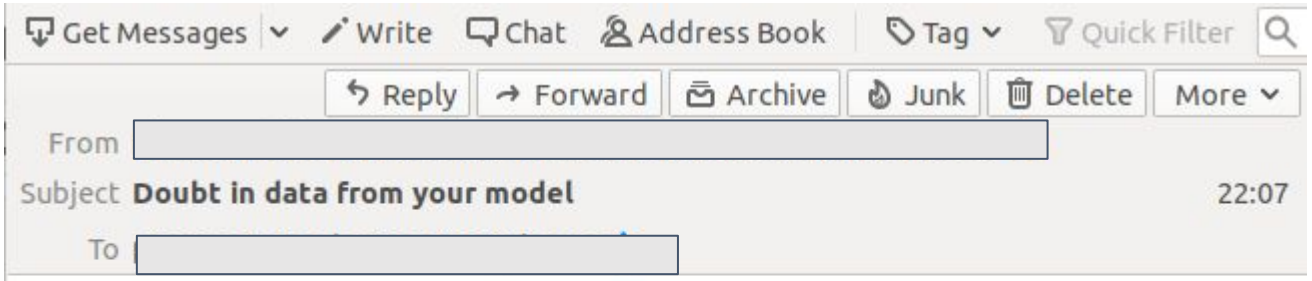
3. What about EC-Earth?
   a. Current status
   b. Roadmap for developments

QA overview: why checking the data?

1. Increase trustability
2. Avoid making wrong scientific decisions
3. Long simulation times imply non rerunnable
4. Avoid having to exchange long emails to clarify data issues:

QA overview: why checking the data?

1. Increase trustability
2. Avoid making wrong scientific decisions
3. Long simulation times imply non rerunnable
4. Avoid having to exchange long emails to clarify data issues:


Get Messages | Write | Chat | Address Book | Tag | Quick Filter

Reply | Forward | Archive | Junk | Delete | More

From [redacted]

Subject **Doubt in data from your model** 22:07

To [redacted]

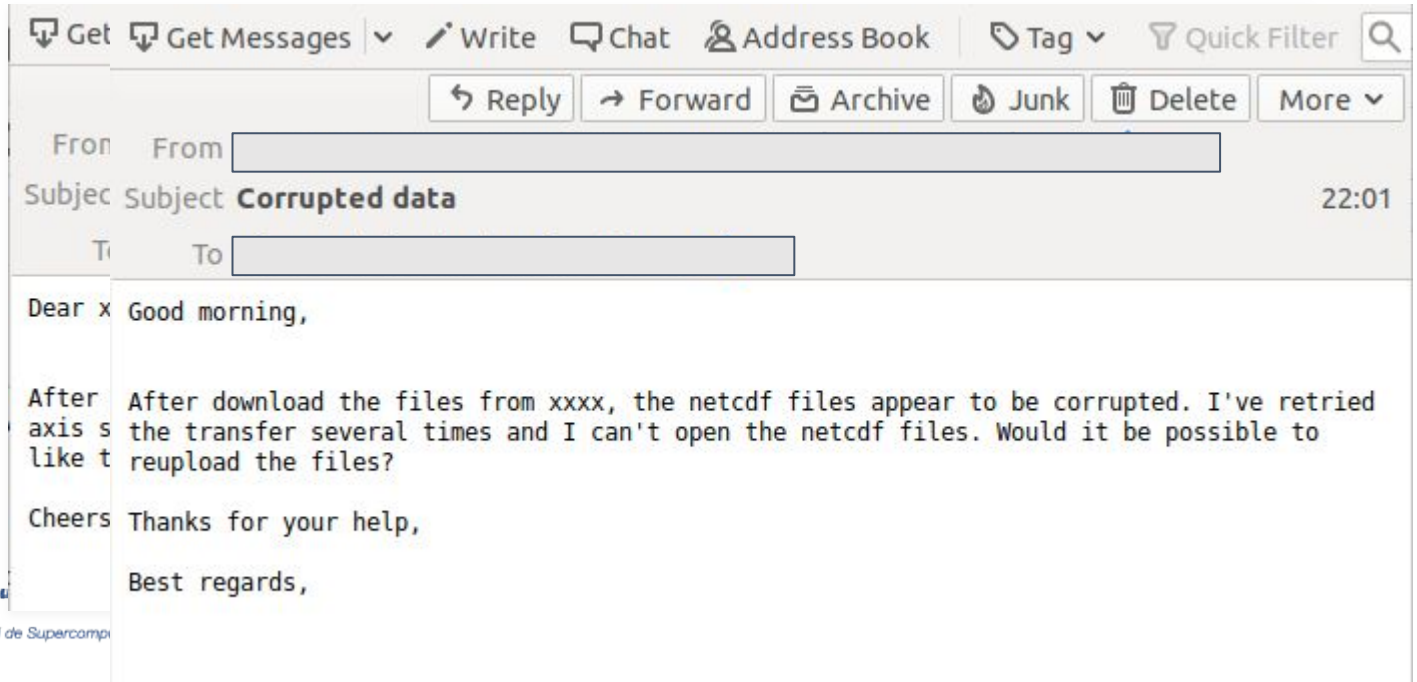
Dear xxxx,

After using the simulations from your experiment called xxxx , it appears that the time axis seems to be shifted one month. Could this be corrected? Or is this supposed to be like this?

Cheers,

QA overview: why checking the data?

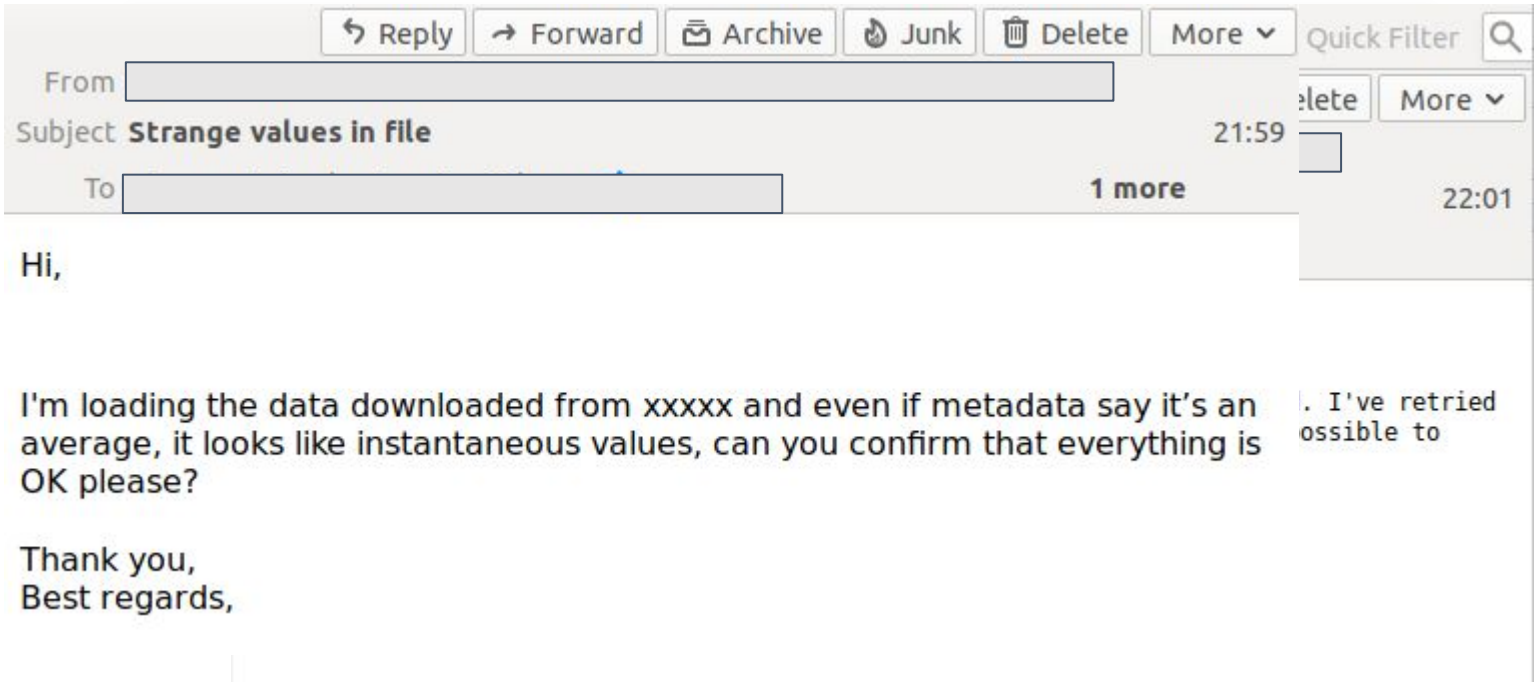
1. Increase trustability
2. Avoid making wrong scientific decisions
3. Long simulation times imply non rerunnable
4. Avoid having to exchange long emails to clarify data issues:



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

QA overview: why checking the data?

1. Increase trustability
2. Avoid making wrong scientific decisions
3. Long simulation times imply non rerunnable
4. Avoid having to exchange long emails to clarify data issues:



1. **Quality Assurance overview**
   a. Why checking the data?
   b. **Potential sources of errors**
   c. What is a good dataset?

2. QA project and software inventory
   a. C3S512
   b. Existing data checkers

3. What about EC-Earth?
   a. Current status
   b. Roadmap for developments

QA overview: sources of “errors”

1. Hardware or system tools failures -> data corruption
   - (s)cp errors, file system failure during writing,...:
     - *ncdump: myfile.nc: NetCDF: Unknown file format*
     - *GRIB_API ERROR : unable to create index for input file ICMGGa1i2+196001 (Wrong message length)*

2. Workflow errors -> the data is not what we expect it to be
   - cmor done before leg+1 resulting in missing timestep
   - rerun on existing files -> extra timesteps in grib

3. Software bugs -> we don’t produce what we think we do
   - date in the files is not what we expect

QA overview: sources of “errors”

4. User configuration mistakes -> we don’t produce what we are supposed to
    - wrong forcing files are read
    - wrong variables/frequencies in varlists

5. “Scientific configuration” of the model/experiment
    - pptd is not well written
    - mask is not read correctly

1. **Quality Assurance overview**
   a. Why checking the data?
   b. Sources of errors
   c. **What is a good dataset?**

2. QA project and software inventory
   a. C3S512
   b. Existing data checkers

3. What about EC-Earth?
   a. Current status
   b. Roadmap for developments

What is a good dataset?

File integrity

Metadata conformance

Missing values

Dates validity

Valid ranges

Spread of the members

Scientific assessment

Informatics

Climate sciences

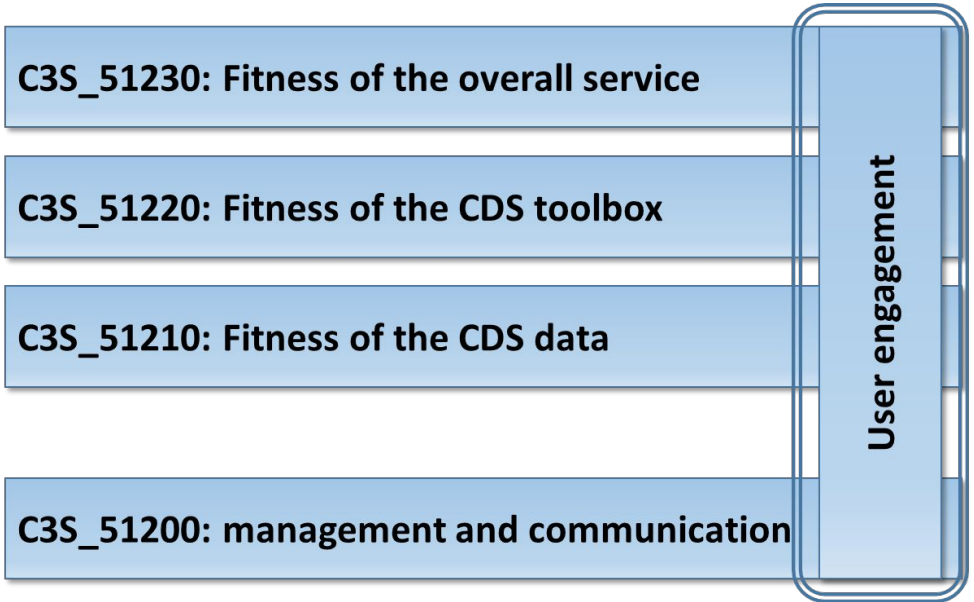
1. Quality Assurance overview
   a. Why checking the data?
   b. Sources of errors
   c. What is a good dataset?

2. **QA project and software inventory**
   a. **C3S512**
   b. Existing data checkers

3. What about EC-Earth?
   a. Current status
   b. Roadmap for developments

C3S_512: QA for the Climate Data Store

- Objective:
  - Providing a user-led overarching EQC service for the whole CDS
  - Providing an independent quality assessment for a number of data types (seasonal forecasts, climate projections and in-situ observations)
- Participants:
  - BSC, DWD, FMI, KNMI, Météo-France, Predictia, CNR, WENR



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

C3S_512: QA for the Climate Data Store

https://cds.climate.copernicus.eu/cdsapp#!/dataset/projections-cmip5-monthly-pressure-levels?tab=form

Please note that some ERA5 downloads might occassionally fail due to internal system issues. Please contact user support for further information.

Overview | Download data | Documentation

Variable

At least one selection must be made

- Temperature
- V-component of wind
- Relative humidity
- U-component of wind
- Specific humidity
- Geopotential height

Select all

Model

At least one selection must be made

- inmcm4 (INM, Russia)
- bcc-csm1-1 (BCC, China)
- BNU-ESM (BNU, China)
- CNRM-CM5 (CNRM-CERFACS, France)
- GFDL-ESM2G (NOAA, USA)
- GISS-E2-H (NASA, USA)
- GISS-E2-R (NASA, USA)
- HadGEM2-CC (UK Met Office, UK)
- IPSL-CM5A-LR (IPSL, France)
- IPSL-CM5B-LR (IPSL, France)
- MPI-ESM-MR (MPI, Germany)
- ACCESS1-0 (BoM-CSIRO, Australia)
- bcc-csm1-1-m (BCC, China)
- CMCC-CMS (CMCC. Italy)
- GFDL-CM3 (NOAA, USA)
- GFDL-ESM2M (NOAA, USA)
- GISS-E2-H-CC (NASA, USA)
- GISS-E2-R-CC (NASA, USA)
- HadGEM2-ES (UK Met Office, UK)
- IPSL-CM5A-MR (IPSL, France)
- MPI-ESM-LR (MPI, Germany)
- NorESM1-M (NCC, Norway)

Select all

Experiment

At least one selection must be made

- AMIP
- Historical
- Pi-control
- RCP 2.6
- RCP 4.5
- RCP 6.0
- RCP 8.5

Select all

Ensemble member

Contact

copernicus-support@ecmwf.int

License

CMIP5 - Data Access - Terms of Use

Publication Date

2016-06-14

Related data

CMIP5 daily data on pressure levels

CMIP5 monthly data on single levels

CMIP5 daily data on single levels

C3S_512: QA for the Climate Data Store

- Main QA tasks: for each data type
  - develop a data checker checking the minimum requirements:
    - file integrity
    - grib keys
    - valid ranges
    - ensemble spread
    - time completeness
  - develop softwares able to do automatic scientific assessment of the data computing different metrics:
    - skill scores
    - RMSE
    - FairRPSS
    - ….

1. Quality Assurance overview
   a. Why checking the data?
   b. Sources of errors
   c. What is a good dataset?

2. **QA project and software inventory**
   a. C3S512
   b. Existing data checkers

3. What about EC-Earth?
   a. Current status
   b. Roadmap for developments

Existing data checkers

- CF checker:
  - **Developed by:** CEDA
  - **Run in:** command line and “drag and drop”
  - **Type of file checked:** netcdf
  - **Checks:** units, standard/long names, dimensions
  - **Reference:** CF conventions (html)
  - **Link(s):**
    - http://pumatest.nerc.ac.uk/cgi-bin/cf-checker.pl
    - https://github.com/cedadev/cf-checker

Existing data checkers

- PrePARE:
  - **Developed by:** PCMDI
  - **Run in:** command line
  - **Type of file checked:** CMIP6 netCDF
  - **Checks:** file names and metadata
  - **Reference:** MIP tables
  - **Link(s):**
    - https://cmor.llnl.gov/mydoc_cmip6_validator/

Existing data checkers

- ESMValTool
  - **Developed by:** DLR, AWI, BSC, NLeSC, Ludwig Maximilian University of Munich, University of Reading
  - **Run in:** through python jobs
  - **Type of file checked:** netcdf “CMOR-like”
  - **Checks:** metadata compliance
  - **Reference:** MIP tables
  - **Link(s):**
    - https://github.com/ESMValGroup/ESMValTool

Existing data checkers

- UKCP18-CC
  - **Developed by:** CEDA
  - **Run in:** command line
  - **Type of file checked:** netcdf
  - **Checks:** file and directory structure, metadata compliance,...
  - **Reference:** rules from user defined json files
  - **Link(s):**
    - https://github.com/ukcp-data/cc-plugin-ukcp18

Existing data checkers

- nctime:
  - **Developed by:** IPSL
  - **Run in:** command line
  - **Type of file checked:** netcdf CMIP6 CMOR
  - **Checks:** time completion of set of files, time “validity”
  - **Reference:** N/A
  - **Link(s):**
    - https://github.com/Prodiguer/nctime

Existing data checkers

- C3S512 grib checker (under development):
  - **Developed by:** BSC
  - **Run in:** command line
  - **Type of file checked:** CDS grib files (model and reanalysis)
  - **Checks:** grib keys, ensemble spread, valid ranges,
  - **Reference:** grib tables, MIP tables
  - **Link(s):**
    - https://earth.bsc.es/gitlab/ces/c3s512-wp1-datachecker (private)

Existing data checkers

- PRIMAVERA checker
  - **Developed by:** CEDA
  - **Run in:** command line
  - **Type of file checked:** netCDF CMOR PRIMAVERA
  - **Checks:** metadata, time coherency, file integrity
  - **Reference:** MIP tables,...
  - **Link(s):**
    - https://github.com/PRIMAVERA-H2020/primavera-val

Existing data checkers

- and more:
    - CMOR
    - check-nemo-files (Klaus Zimmermann)
    - QA-DKRZ

1. Quality Assurance overview
   a. Why checking the data?
   b. Sources of errors
   c. What is a good dataset?

2. QA project and software inventory
   a. C3S512
   b. Existing data checkers

3. **What about EC-Earth?**
   a. **Current status**
   b. Roadmap for developments

What about EC-Earth?

- What do we (currently, operationally, automatically) check?

What about EC-Earth?

- What do we (currently, operationally, automatically) check?
  - at the consortium level: (almost) NOTHING!

What about EC-Earth?

- What do we (currently, operationally and automatically) check?
  - at the institute level:
    - CMORization/ece2cmor (all, online or offline)
    - “some checks” for NaN on Nemo files (SMHI)
    - “some checks” on CMOR files (CNR)
    - “some checks” on grib files (BSC)
    - “more extensive checks” on pre-ESGF publication (BSC)
    - … more?

1. Quality Assurance overview
   a. Why checking the data?
   b. Sources of errors
   c. What is a good dataset?

2. QA project and software inventory
   a. C3S512
   b. Existing data checkers

3. **What about EC-Earth?**
   a. Current status
   b. **Roadmap for developments**

What about EC-Earth?

- Check “original/raw” and CMOR files

File integrity → Metadata conformance → Missing values → Dates validity → Valid ranges → Spread of the members → Scientific assessment

Informatics — Climate sciences

What about EC-Earth?

- Check “original/raw” and CMOR files
  - grib_dump
  - cdo info

✓ ✓

File integrity | Metadata conformance | Missing values | Dates validity | Valid ranges | Spread of the members | Scientific assessment

Informatics → Climate sciences

What about EC-Earth?

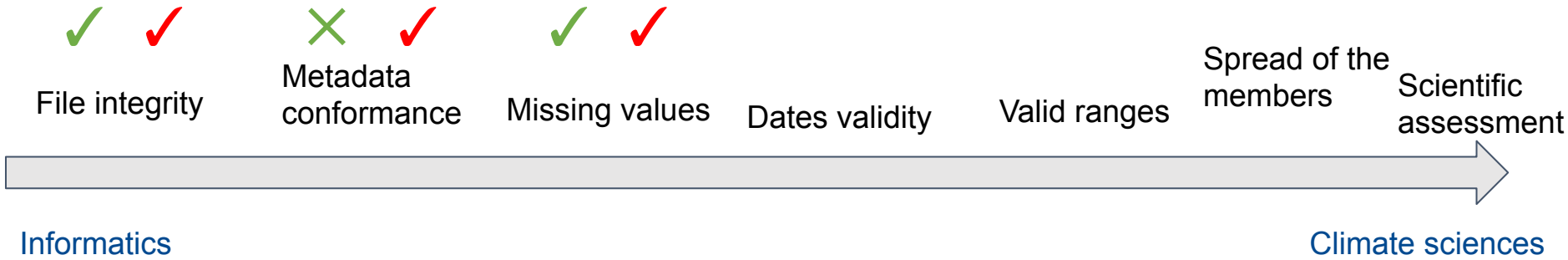
- Check “original/raw” and CMOR files
  - ece2cmor/CMOR
  - PREPARE
  - UKCP18-CC

File integrity	Metadata conformance	Missing values	Dates validity	Valid ranges	Spread of the members	Scientific assessment
✓ ✓	✗ ✓					

Informatics → Climate sciences

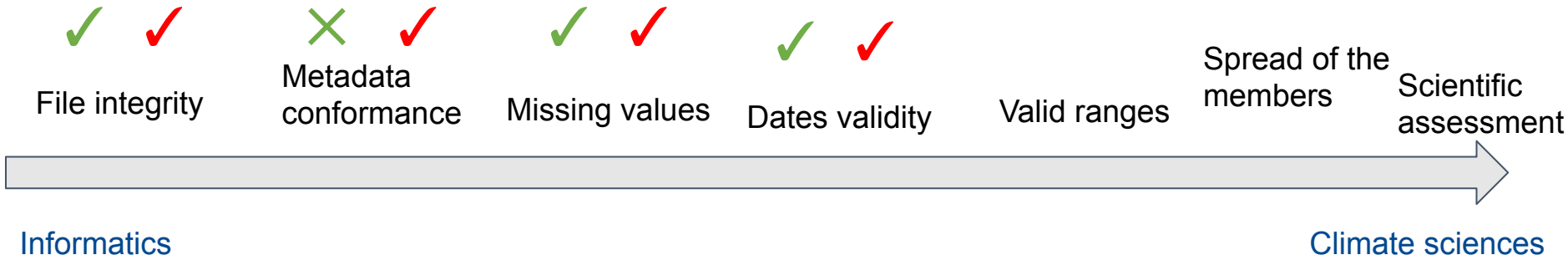
What about EC-Earth?

- Check “original/raw” and CMOR files
  - cdo/nco
  - check_nemo_files



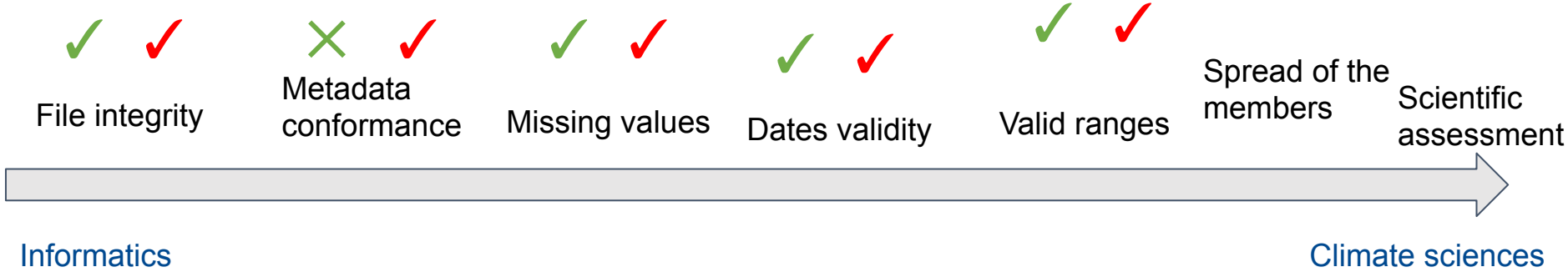
What about EC-Earth?

- Check “original/raw” and CMOR files
  - ???
  - nctime



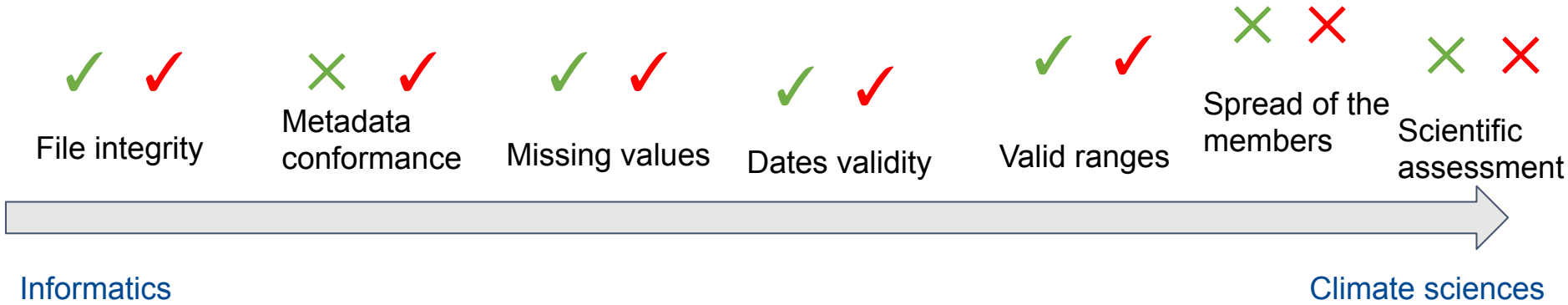
What about EC-Earth?

- Check “original/raw” and CMOR files
  - C3S_512 grib_checker
  - ece2cmor/CMOR



What about EC-Earth?

- Check “original/raw” and CMOR files



Conclusions

- It is impossible to QA everything automatically but there is **room for improvement**

- Possible checks to be implemented:
  - **file corruption** for raw IFS and Nemo
  - **timesteps/date coherence** for raw IFS and Nemo and CMOR files
  - **masks and NaN** for Nemo raw files
  - **variable list completeness** for CMOR files
  - **metadata coherency** for CMOR files (PrePARE)
  - **time completeness** for CMOR files (nctime)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thank you

pierre-antoine.bretonniere@bsc.es