Towards an ORCHIDEE evaluation chain

Patricia CADULE[1] Sebastián RIFFO[2]

Institut Pierre-Simon Laplace (IPSL), CNRS[1] and Sorbonne Université[2]

Monday 10th February, 2025

Context

Task force (2020-22)

Context

Task force (2020-22)

- Methodology (variables, gathering data, scripting)

Context

Task force (2020-22)

- Methodology (variables, gathering data, scripting)
- Hands-on approach

Context

Task force (2020-22)

- Methodology (variables, gathering data, scripting)
- Hands-on approach
- Visualization expectations

Context

Task force (2020-22)

- Methodology (variables, gathering data, scripting)
- Hands-on approach
- Visualization expectations

Accomplishments

- Evaluation dataset (at February 2025: 13/24 variables)

Context

Task force (2020-22)

- Methodology (variables, gathering data, scripting)
- Hands-on approach
- Visualization expectations

Accomplishments

- Evaluation dataset (at February 2025: 13/24 variables)

Challenges

- Setting unambiguous and objective criteria for evaluation

Context

Task force (2020-22)

- Methodology (variables, gathering data, scripting)
- Hands-on approach
- Visualization expectations

Accomplishments

- Evaluation dataset (at February 2025: 13/24 variables)

Challenges

- Setting unambiguous and objective criteria for evaluation
- Having a reliable dataset to contrast against

Context

Task force (2020-22)

- Methodology (variables, gathering data, scripting)
- Hands-on approach
- Visualization expectations

Accomplishments

- Evaluation dataset (at February 2025: 13/24 variables)

Challenges

- Setting unambiguous and objective criteria for evaluation
- Having a reliable dataset to contrast against
- Development of a open, intuitive, and robust tool

Evaluation chain

Evaluation framework

Quality assessment

Data considerations

Evaluation chain

Evaluation framework

Against what we contrast and why?

Quality assessment

How do we define *being better*?

Data considerations

Setting standards for its intended use

Evaluation framework

The *notion of model evaluation* varies from team to team, in accordance to their own requirements and criteria

Evaluation framework

The *notion of model evaluation* varies from team to team, in accordance to their own requirements and criteria

PLUMBER Protocol (Abramowitz et al)

- *evaluation*
  determine errors with respect to observations, using statistical methods
- *comparison*
  viewing a model in relation to others, with the purpose of establishing differences in their behavior
- *benchmarking*
  identifying, from the contrast with a reference (defined in advance), what a model requires to improve

Visualization

Useful to convey information to a wide public, yet an explanation of these results relies in the observer's experience, rather than a *settled criterion*

Quality assessment

Visualization

Useful to convey information to a wide public, yet an explanation of these results relies in the observer's experience, rather than a *settled criterion*

Quantification

Delivering a *measurable* standard depends on a shared space-time domain

- *metrics*
  measures (statistics), indicators (process-oriented), scores (as ILAMB)

Quality assessment

Visualization

Useful to convey information to a wide public, yet an explanation of these results relies in the observer's experience, rather than a *settled criterion*

Quantification

Delivering a *measurable* standard depends on a shared space-time domain

- *metrics*
  measures (statistics), indicators (process-oriented), scores (as ILAMB)

Decision-making then based on

- *orders of magnitude*
- *sensitivity*

to recognize relevant variables in the model's behavior

Data considerations

Data's nature and its processing have and *overlooked impact* in model quality assessment

Data considerations

Data's nature and its processing have and *overlooked impact* in model quality assessment

Operational dataset

Setting standards around

- *scale*
  original resolution, spatial and temporal coverage
- *regridding*
  upscaling (interpolations) or downscaling (e.g. averaging) effects
- *reliability*
  accounts for variability and uncertainty, using an ensemble of realizations or sources

Evaluation chain

Evaluation framework

- evaluation
- comparison
- benchmarking

Quality assessment

- visualization
- metrics
- orders of magnitude
- sensitivity

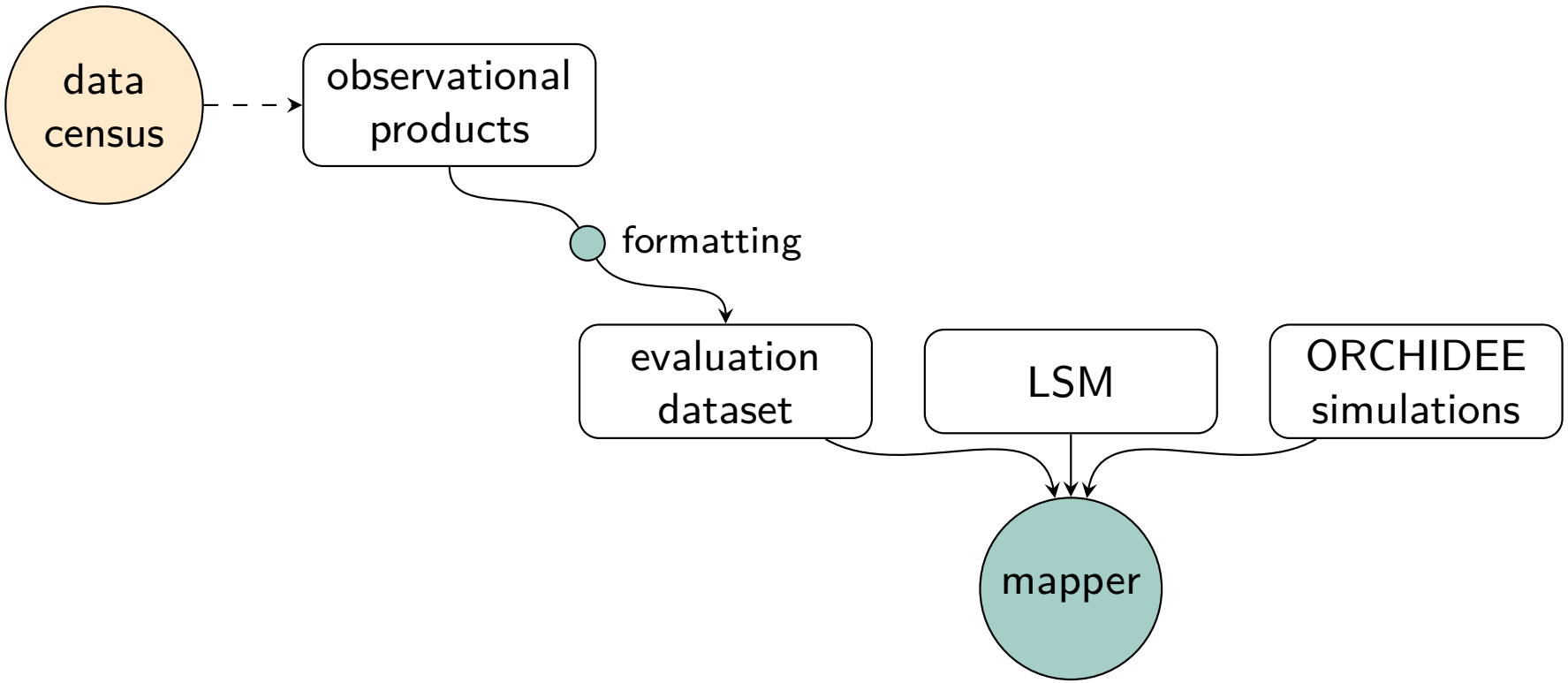
Data considerations

- scale
- regridding
- reliability

Evaluation tool

> *The tool should « be open (...), documented enough (...), easy [exporting the figures for possible publication]; not require (...) to have specific knowledge of where datasets are hosted (...); provide a basic set of typical evaluation techniques (...) »*
>
> Task Force (2021)

Evaluation tool



Evaluation tool

MOMA (Metrics for ORCHIDEE Model Analysis)

- Web application running on spirit
- Interactive and dynamic visualization
- Real-time catalog
- Statistical measures

Evaluation tool

MOMA (Metrics for ORCHIDEE Model Analysis)

- Web application running on spirit
- Interactive and dynamic visualization
- Real-time catalog
- Statistical measures

Technical stack

- Front end
  JavaScript (D3, Gridstack, DataTables)
- Back end
  Python (Flask; Xarray, Numpy, Dask)
- Version control
  Git

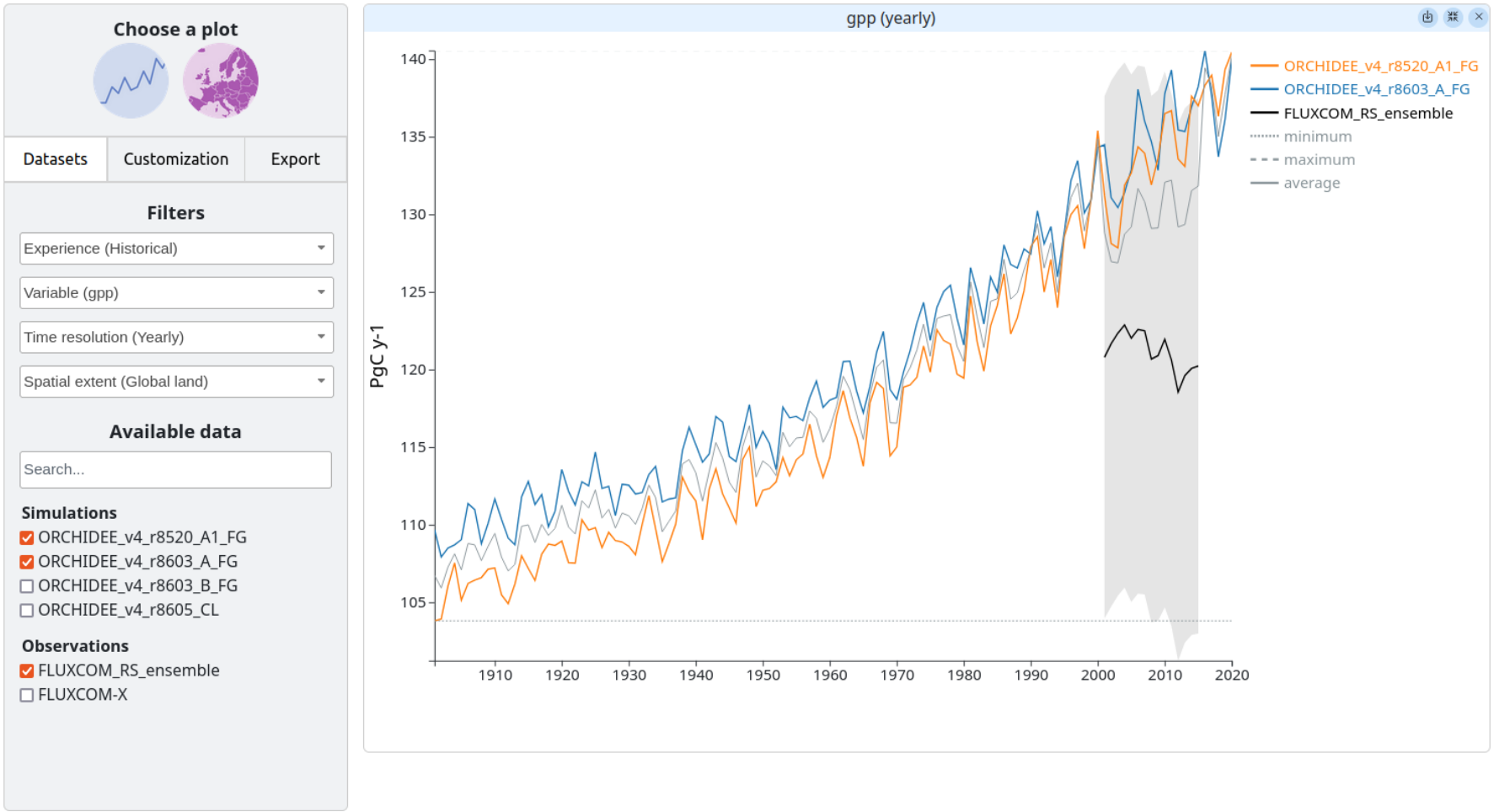


Figure 1: Front end

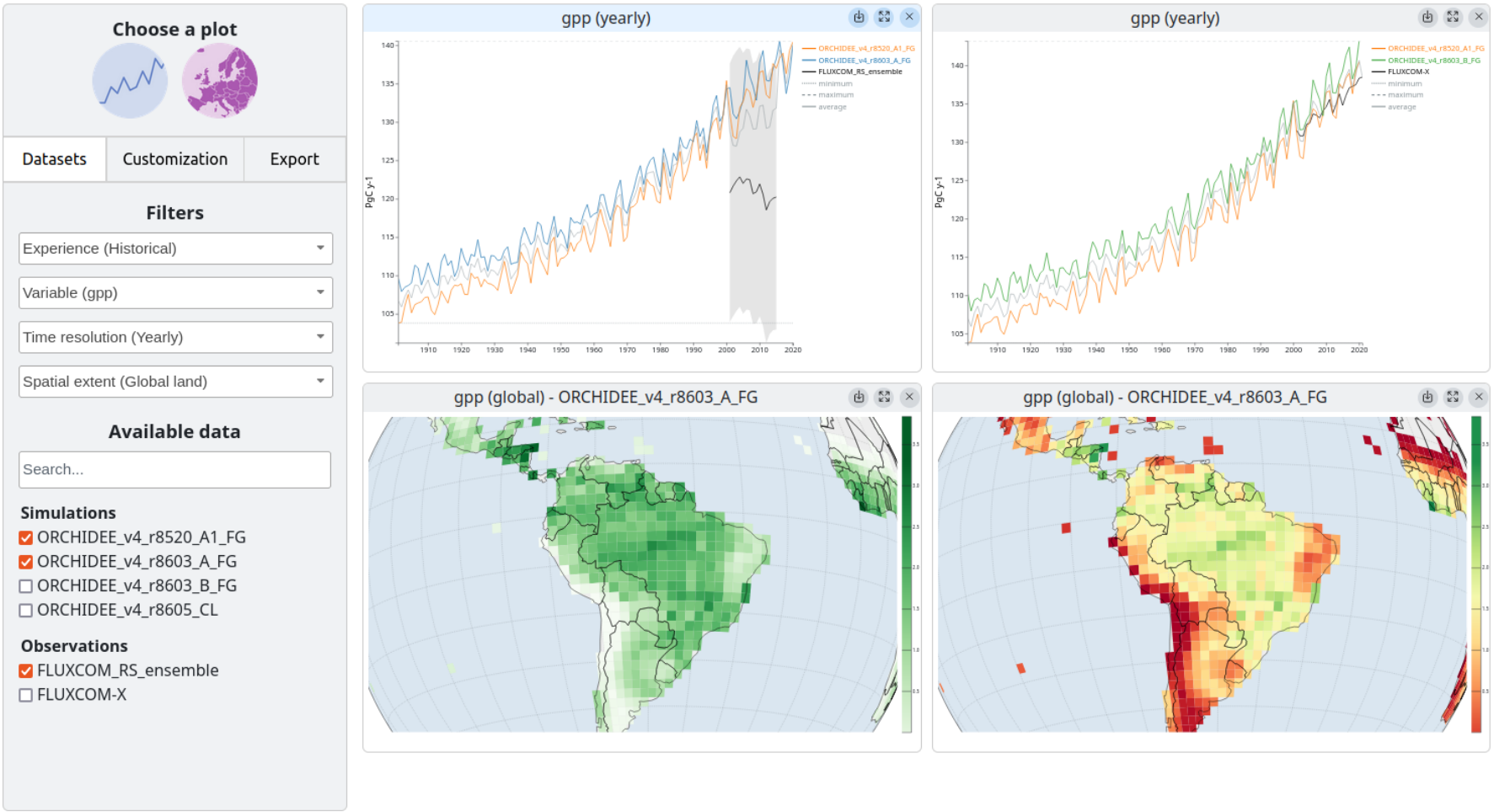
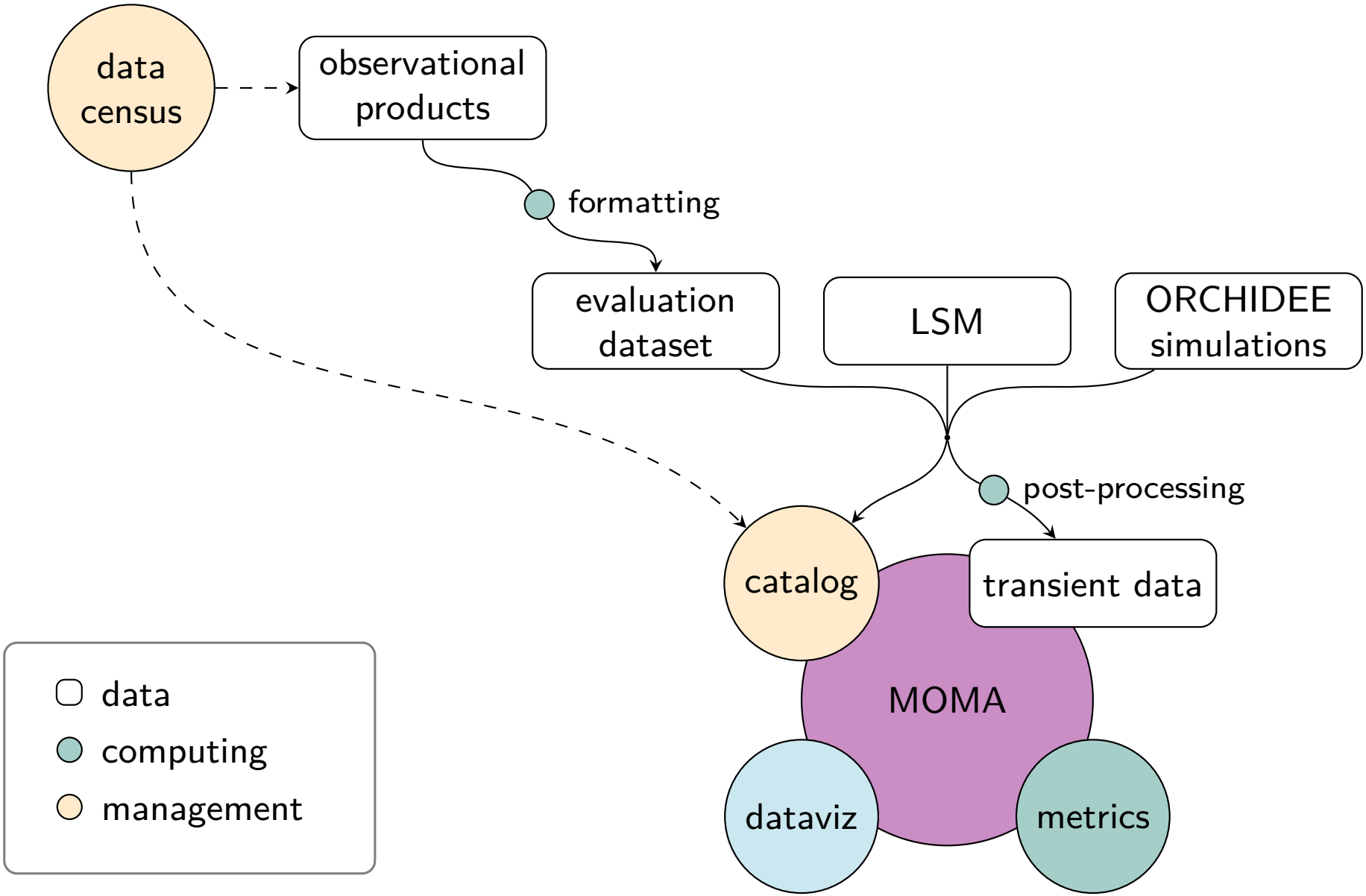
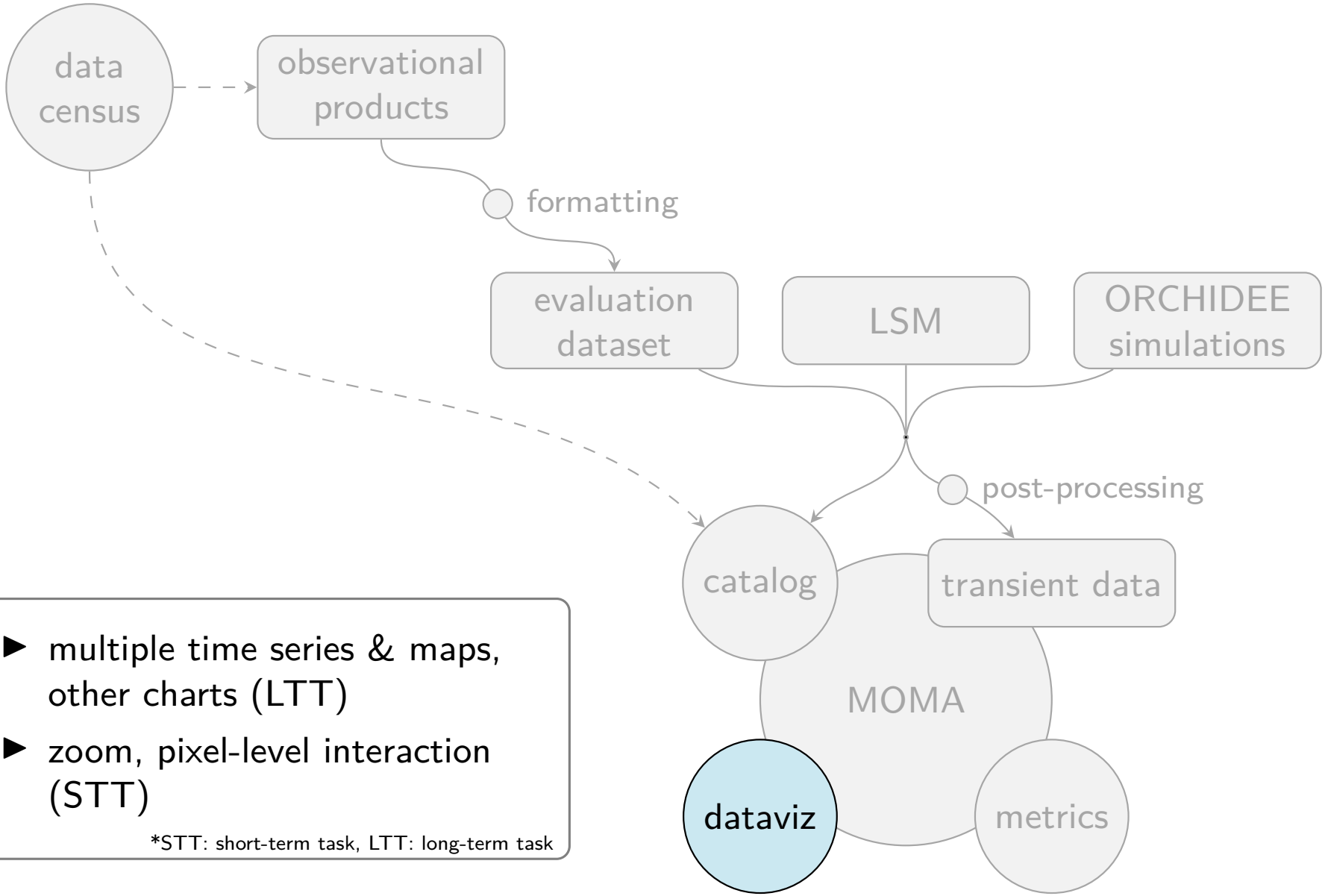


Figure 2: Front end

Metrics for ORCHIDEE Model Analysis



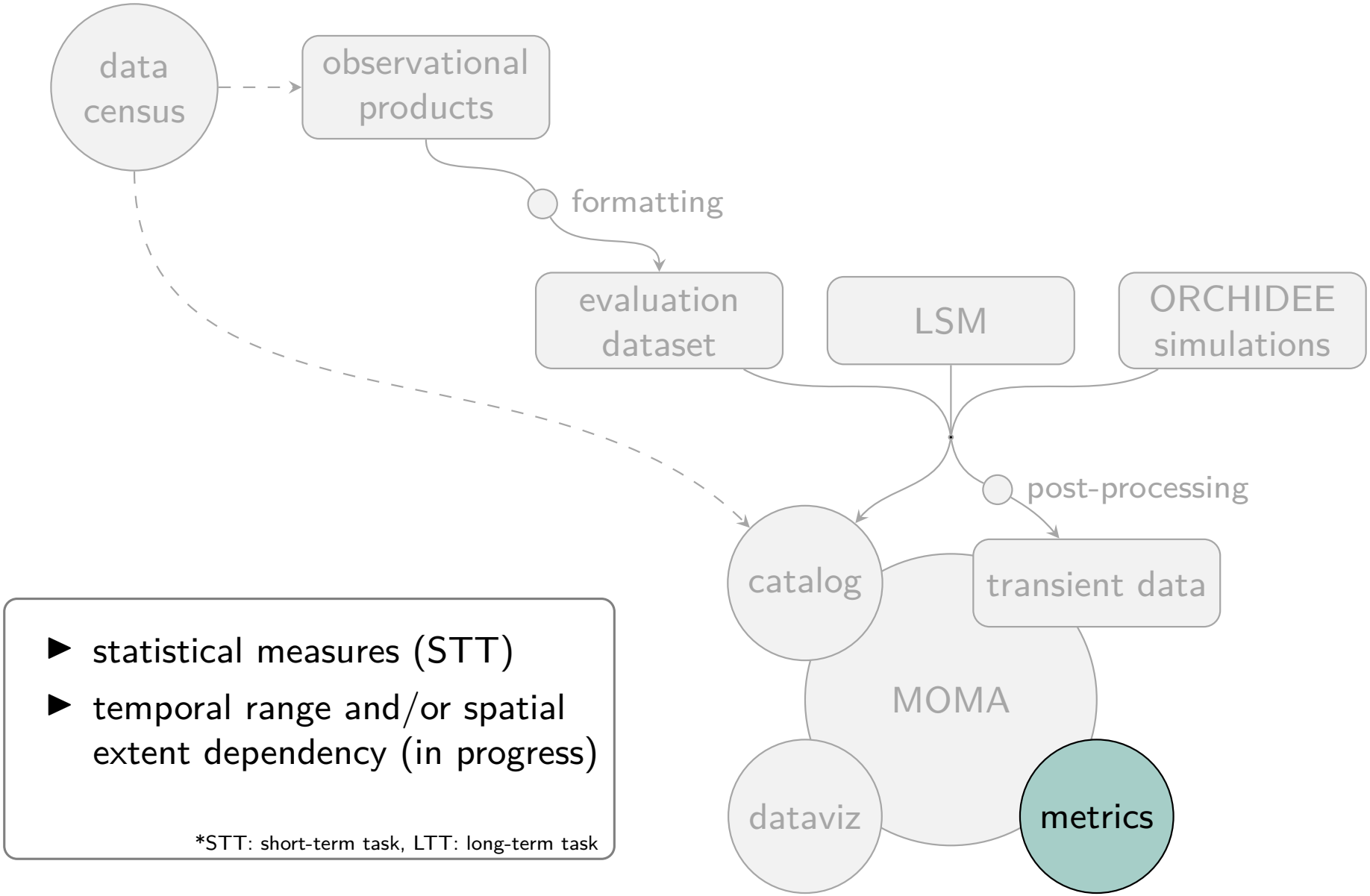
Metrics for ORCHIDEE Model Analysis



- multiple time series & maps, other charts (LTT)
- zoom, pixel-level interaction (STT)

*STT: short-term task, LTT: long-term task

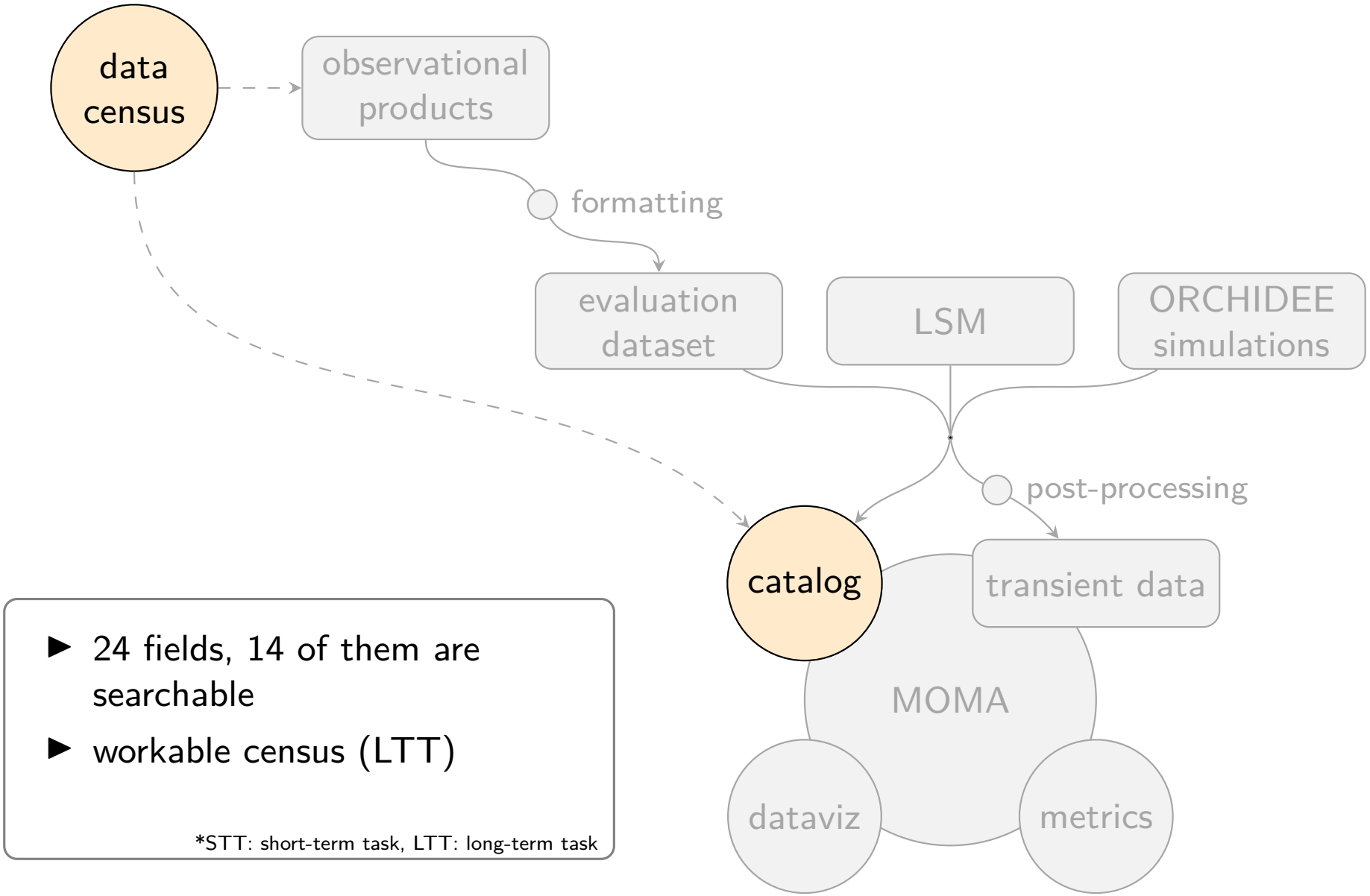
Metrics for ORCHIDEE Model Analysis



- statistical measures (STT)
- temporal range and/or spatial extent dependency (in progress)

*STT: short-term task, LTT: long-term task

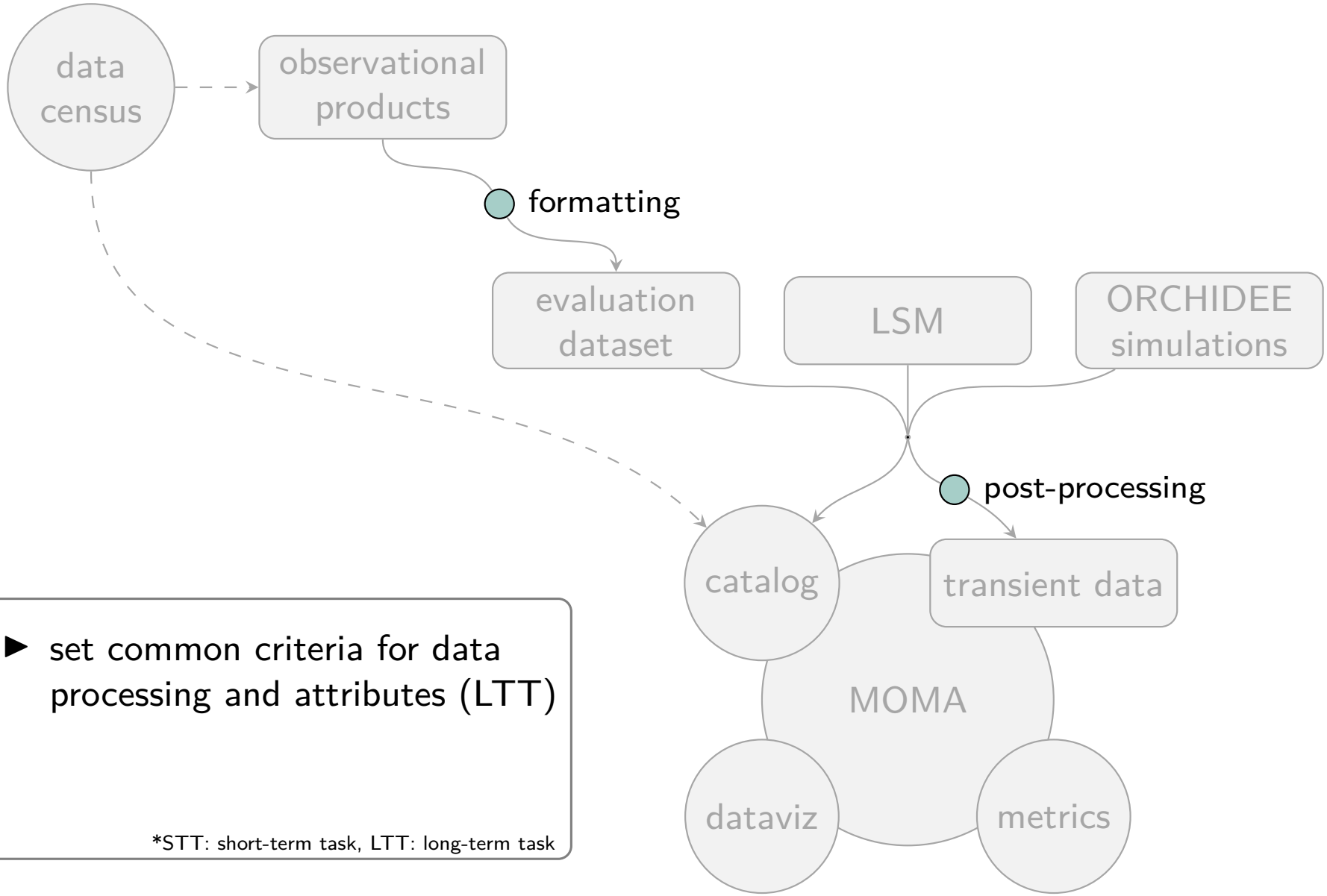
Metrics for ORCHIDEE Model Analysis



- 24 fields, 14 of them are searchable
- workable census (LTT)

*STT: short-term task, LTT: long-term task

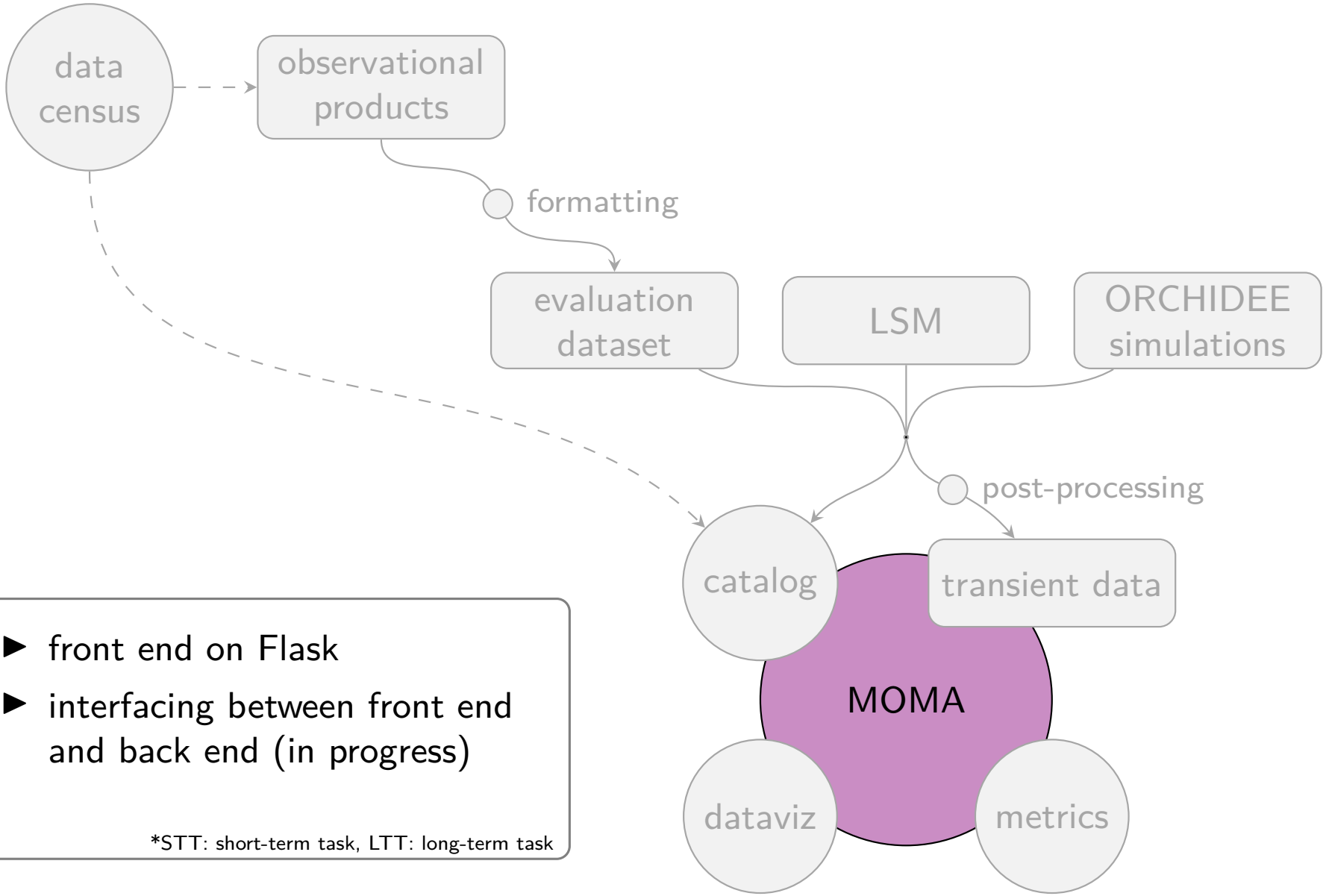
Metrics for ORCHIDEE Model Analysis



- set common criteria for data processing and attributes (LTT)

*STT: short-term task, LTT: long-term task

Metrics for ORCHIDEE Model Analysis



- front end on Flask
- interfacing between front end and back end (in progress)

*STT: short-term task, LTT: long-term task

Other evaluation tools

Nationwide

- CliMAF & C-ESM-EP (CNRM and IPSL)
- ILAMB (RUBISCO, US)
- AMBER (CLASSIC, Canada)
- PMP (LLNL, US)
- LVT (NASA, US)

International

- ESMValTool
- PLUMBER2 MIP

Communication

Wiki

https://forge.ipsl.jussieu.fr/orchidee/wiki/Documentation/Validation

Report (in progress)

- Documentation
- Publication

Thank you for your attention !