NFDI for seamless Earth system model-data integration

Matthias Forkel[1], Kirsten Thonicke[2], and Nuno Carvalhais[3]

[1]Technische Universität Dresden, Junior professorship for Environmental Remote Sensing
[2]Potsdam Institute for Climate Impact Research, Earth System Analysis
[3]Max Planck Institute for Biogeochemistry, Department for Biogeochemical Integration

23.07.2020

Abstract

Global Earth observation data is invaluable to evaluate, parametrize and further develop Earth system models and its land components, dynamic global vegetation models. However, such an integration of Earth observation data with Earth system models requires the expertise of multiple institutions, and in practical terms often involves converting file formats, copying large datasets between institutions and related computing systems, and manual and time-consuming application of multiple scripts and program code. The objective of the pilot is to identify requirements and to define a generic framework for a seamless integration of Earth observation data with Earth system models. Specifically, we will outline and develop a prototype for a seamless workflow to apply satellite observations for benchmarking and parametrization of the LPJmL dynamic global vegetation model as part of the Potsdam Earth Model (POEM). The expected results will enable a comprehensive and continuous use of satellite observations for the development of LPJmL and POEM. Moreover, the developed framework and prototype will guide the development of seamless infrastructures to integrate Earth observations and models in the global biogeochemical, hydrological, ecological, and climate science communities and hence in NFDI4Earth as a whole.

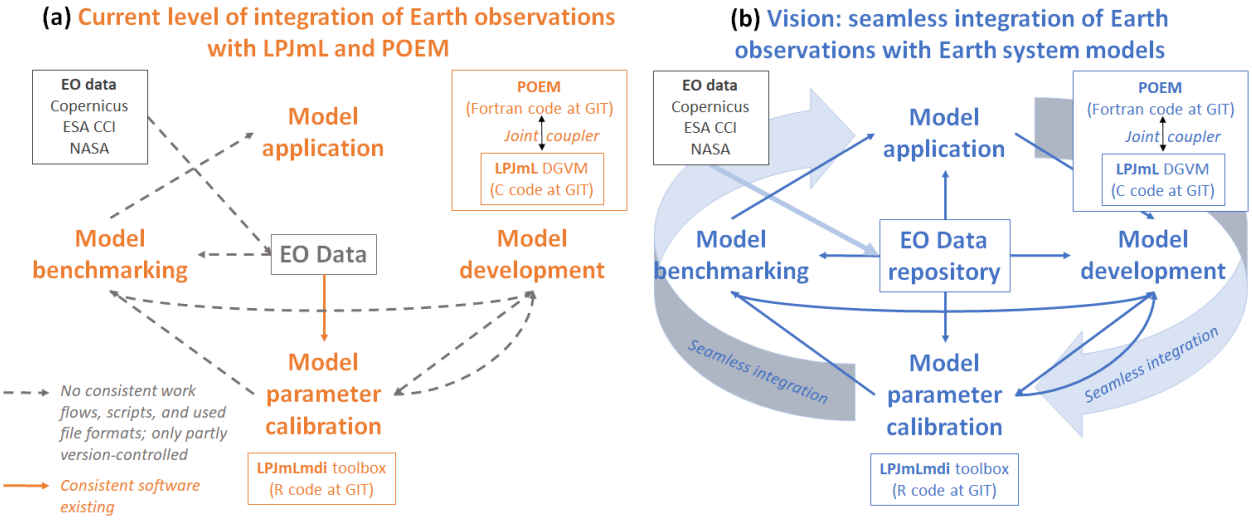

Figure 1: Current state vs. seamless integration of Earth observations with Earth system models and global vegetation models.

I. Introduction

Earth system models (ESMs) are the main tools to estimate past and future developments of the interactions between atmosphere, ocean and land. However, model predictions of e.g. future temperature changes are uncertain because of a limited understanding of the underlying processes or because of model parametrizations. Earth observations (EO) from satellites can be used to benchmark ESMs, guide model development, calibrate model parameters and hence to potentially reduce these uncertainties[1,2]. This is especially the case for the land components of ESMs, so called dynamic global vegetation models (DGVMs)[4,5]. For example, EO products were used to parametrize modeled phenology[6], photosynthesis[7], fire dynamics[8,9], or general vegetation dynamics and carbon turnover[2] in DGVMs. The use of EO data for the development and parametrization of DGVMs is now starting to propagate in the development of ESMs. For example, new modules were developed in the DGVM LPJmL based on EO data[8,10] and are now included in the Potsdam Earth Model (POEM)[11]. A continuous **model-data integration (MDI)** cycle involves model development, parameter calibration, benchmarking, application and potentially model reformulation[12]. Model calibration and model evaluation tools are at the core of MDI and have been implemented for LPJmL based on previous research[2,6,8] in the LPJmLmdi software package[13].

However, the integration of EO data with ESMs or DGVMs currently lacks a formal infrastructure for seamless integration (Figure 1) but involves the use of different file formats (e.g. binary, tiff, NetCDF), of various scripts and tools for EO data processing (e.g. written in Shell, Python, R or CDO), and of different scripts or packages for model benchmarking, parameter calibration (e.g. LPJmLmdi) and application (in various languages). Although LPJmL, POEM, and LPJmLmdi are hosted in GIT repositories, most application scripts are not version-controlled and nobody has a complete overview about all datasets, scripts or packages that were used for model-data integration work. Additionally, often multiple researchers, institutions and hence computing facilities are involved in model-data integration which practically implies copying large datasets between IT infrastructures. The lack of a formal infrastructure and of easy to use software packages currently prevents other modeling groups to adopt MDI approaches and hence hampers harvesting the full potential of EO data and model-data integration approaches for the development of DGVMs and ESMs.

Our vision is an infrastructure that allows a seamless integration of EO with ESMs and DGVMs (N4E-MDI). Such an infrastructure would allow a continuous, fast and productive use of EO data in model development, calibration, benchmarking and application without hurdles of data conversion, data copying, code inconsistencies, and a high level of manual technical intervention. As a result, the infrastructure will advance the development of global models. We aim to develop a generic **N4E-MDI framework** for integrating EO data with ESMs and DGVMs and will implement a first **N4E-MDI prototype** around LPJmL, POEM and the LPJmLmdi package.

II. Pilot description

The development of the N4E-MDI framework and of the prototype will be conducted in four work packages. We will assess requirements for a framework at ESM and DGVM groups through specialized interviews (*WP1*). Based on these requirements, an overall generic N4E-MDI framework will be defined (*WP2*). Based on the framework, a N4E-MDI prototype will be

implemented (*WP3*) and tested in a demonstration case (*WP4*). The results from all work packages will be synthesized in a final **roadmap report**. According to our current understanding, the N4E-MDI framework and the associated prototype will likely consist of several components to cover all aspects of model-data integration:

- **EO data repository:** All EO datasets will be saved as NetCDF4 climate convention files to allow an easy uptake and access by the Earth system community. Within the prototype a simple folder structure will serve as data repository. However, we envision that an open cloud-based data cube should serve as data repository in the long-term. We expect that N4E-MDI will benefit from other developments within NFDI4Earth.
- **EO data reader toolbox:** Data readers will process and convert various EO datasets into the required format for the data repository. We aim to implement automatic work flows that ensure a frequent update of some common EO datasets. For the prototype, we will implement these workflows for MODIS products and for some datasets of the ESA Climate Change Initiative.
- **Model evaluation toolbox:** This toolbox will contain functions to compute model performance metrics and to make typical plots for model evaluation (e.g. residual plots, Taylor diagrams).
- **Model calibration toolbox:** The model calibration toolbox is based on various optimization algorithms and relies on the performance metrics from the evaluation toolbox to calibrate model parameters. The toolbox will be partly based on the LPJmLmdi package and on recent developments in the group of N. Carvalhais.
- **Model application toolbox:** Here we plan to collect and harmonize frequently used scripts to make standard analyses of model results and related figures (maps and time series).
- **Model-data interface toolbox:** The model interface should enable the exchange of data between the EO data repository, the toolboxes and the ESM or DGVM and should invoke model runs. We aim to implement a generic model interface that can be easily adapted for several models. The prototype will be partly based on the LPJmLmdi package.

All toolboxes will be consistently and tidy implemented in commonly used scientific computing languages (Python and/or R) and will be made available at an open GIT repository. Additionally, we aim to develop easy to use Jupyter notebooks that demonstrate the application of the individual toolboxes, repository and of the overall N4E-MDI framework.

III. Relevance for the NFDI4Earth

We envision that the proposed N4E-MDI framework will be taken up by several ESM and DGVM modelling centers across Germany and beyond. Although the development of the proposed framework is driven by the land modelling and remote sensing community, similar challenges are faced in the ocean and atmosphere communities. The proposed framework will make EO data accessible for Earth system modeling groups, interoperable and reusable in different stages of model development. The application of MDI approaches will be not restricted to specialized studies but will become a standard toolset. In addition, publications can refer to the EO data repository and to the toolboxes which will facilitate the reproducibility of research studies. In the long-term, local infrastructure providers will benefit from a harmonized MDI framework without the need to reinvent the wheel.

IV. Deliverables, work plan and requested funding

The proposed pilot has a duration of 18 months and includes three deliverables (Figure 2):

- D1: An **open accessible blog** will document and summarize all developments during N4E-MDI. Blog posts will be published at least after finishing each task.
- D2: **Software packages** of the prototype will be made available at an open **GIT repository**.
- D3: The **roadmap report** will summarize basic concepts and best practices of model-data integration and synthesize the N4E-MDI framework, prototype and the results of the first practical demonstration. The future development of MDI infrastructures within NFDI4Earth will be outlined.

The proposed work will be conducted by one PostDoc (65% FTE, E13 level, 18 months). We request additional funding for short research visits to the partner institutions PIK Potsdam and MPI-BGC Jena to develop the framework and prototype in close cooperation with leading experts and model developers. Hence, we request a **total funding of 75.000 €.**

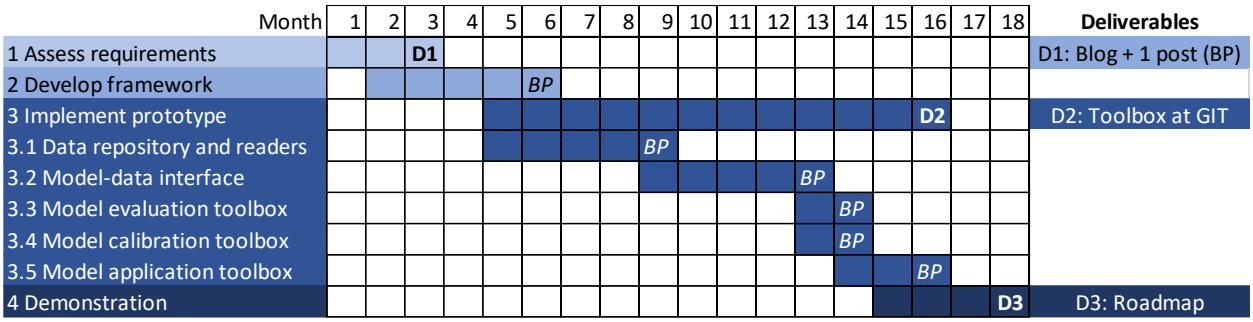

Figure 2: Overview of the work plan including deliverables (D) and blog posts (BP).

References

1. Eyring, V. *et al.* Taking climate model evaluation to the next level. *Nat. Clim. Change* **9**, 102 (2019).
2. Forkel, M. *et al.* Constraining modelled global vegetation dynamics and carbon turnover using multiple satellite observations. *Sci. Rep.* **9**, 1–12 (2019).
3. Forkel, M. *et al.* A data-driven approach to identify controls on global fire activity from satellite and climate observations (SOFIA V1). *Geosci. Model Dev.* **10**, 4443–4476 (2017).
4. Exbrayat, J.-F. *et al.* Understanding the Land Carbon Cycle with Space Data: Current Status and Prospects. *Surv. Geophys.* (2019) doi:10.1007/s10712-019-09506-2.
5. Paula, M. D. de, Giménez, M. G., Niamir, A., Thurner, M. & Hickler, T. Combining European Earth Observation products with Dynamic Global Vegetation Models for estimating Essential Biodiversity Variables. *Int. J. Digit. Earth* **0**, 1–16 (2019).
6. Forkel, M. *et al.* Identifying environmental controls on vegetation greenness phenology through model–data integration. *Biogeosciences* **11**, 7025–7050 (2014).
7. MacBean, N. *et al.* Strong constraint on modelled global carbon uptake using solar-induced chlorophyll fluorescence data. *Sci. Rep.* **8**, 1973 (2018).
8. Drüke, M. *et al.* Improving the LPJmL4-SPITFIRE vegetation–fire model for South America using satellite data. *Geosci. Model Dev.* **12**, 5029–5054 (2019).
9. Forkel, M. *et al.* Emergent relationships with respect to burned area in global satellite observations and fire-enabled vegetation models. *Biogeosciences* **16**, 57–76 (2019).
10. Schaphoff, S. *et al.* LPJmL4 – a dynamic global vegetation model with managed land – Part 1: Model description. *Geosci Model Dev* **11**, 1343–1375 (2018).
11. Drüke, M. *et al. Coupling the dynamic vegetation model LPJmL5.1 to an Earth system model - towards POEM1.0*. https://meetingorganizer.copernicus.org/EGU2020/EGU2020-17982.html (2020) doi:10.5194/egusphere-egu2020-17982.
12. Keenan, TrevorF., Carbone, MariahS., Reichstein, M. & Richardson, AndrewD. The model–data fusion pitfall: assuming certainty in an uncertain world. *Oecologia* **167**, 587–597 (2011).
13. PIK-LPJmL/LPJmLmdi. https://github.com/PIK-LPJmL/LPJmLmdi (2020).