Calibrated Land Model Intercomparison Project 1st Stage Draft Protocol

Background

Motivation

CalLMIP is a new initiative led by members of the Analysis and Integration of Modeling Earth Systems (AIMES) Land Data Assimilation Working Group and the International Land Modeling Forum (ILMF) Parameter Estimation Working Group.

This initiative is targeted at global-scale land models - i.e., land surface or terrestrial biosphere models that form part of Earth system models, or intermediate complexity models and vegetation demographic models that are routinely run at global scale (for example for the global carbon budget).

The overall goals envisaged for CalLMIP are:

- To quantify the reduction in inter-model spread as a result of parameter calibration for a wide range of variables (e.g., carbon and water stores and fluxes, aboveground biomass, soil carbon).
- To encourage the use of uncertainty quantification methods for constraining parameter uncertainty in global land models.
- To better understand and exchange knowledge on the variety of methods being used in global land model parameter calibration in order to identify best practices.

Guiding principles for CalLMIP

1. **First phase goal:** A community-led initiative to evaluate how well models' existing ("in-house") parameter calibration efforts reduce model-data misfit and the spread across models for key variables.

2. **'Come as you are' approach:** To leverage current model parameter estimation systems, rather than requiring development of new systems, maintaining flexibility to accommodate existing methods.

3. **Keep it as simple as possible:** Phase 1 is a starting point from which additional analysis/experiments can develop e.g., no global future projections, no tests of

different calibration set-up.

4. **All outputs available in public repository**: for additional analyses by the community.

5. **Site optimization focus:** Key variables **to opt into** for optimization include carbon fluxes, evapotranspiration/latent heat flux, sensible heat flux (any number of variables to output).

Proposed timeline for the 1st phase of CalLMIP

- March/April 2025: Three community sessions to decide on protocol, data, sites etc.
- May-August 2025: Preparation of input and observation files and data harmonization by the MacBean lab for all modeling groups.
- September 2025: Deliver inputs and finalised protocol to modeling groups. Draft of protocol paper circulated (if we decide to go ahead with this).
- September 2025: Webinar to present the first stage protocol and gather feedback on future CalLMIP phases and goals to the wider community.
- Sept 2025 to March 2026: 1st stage parameter calibration experiments carried out by individual modeling groups.
- March to July 2026: Initial post calibration intercomparison analyses.
- July/Aug/Sept 2026: Presentation of 1st stage results at the 2nd Land Surface Modeling Summit (or equivalent meeting) and online webinar.

Protocol Plan

Red: Still to be decided

1. Calibration experiment set-up

1.1 Elements that will be constant across all experiments and modeling groups

To ensure comparability across modeling groups, certain key aspects of the calibration experiment will be standardized:

- Flux tower data: a common climate forcing dataset will be used for all experiments. This will be prepared by the members of Natasha MacBean's lab group and distributed by the CalLMIP planning committee to participating modeling groups. The dataset will include data for a fixed number of sites (see Section 2) that represent the diversity of PFTs worldwide. For each site, the following will be provided: i) *in situ* meteorological forcing data from the flux tower for the entire calibration and validation period; ii) daily observations of NEE (GPP, Reco), ET/LE, and H, with uncertainties (to be decided – see Section 4), for calibration only; and iii) site meta-data such as soil characteristics and vegetation type.
- Spin-up protocol: each group must ensure that models start from a stable, physically realistic state, preventing biases from arbitrary initial conditions; therefore, a consistent spin-up protocol will be described in this protocol (see Section 7) (unless the group has a set functionality to prescribe initial conditions/states).
- Posterior uncertainties on parameters and variables optimized will be required.
- Output file format will be standardized across models (see Section 6).
- Calibration/validation setup: a set number of sites/years will be set aside for validation (Section 8). This will be determined by the CalLMIP planning committee ahead of the calibration phase of the MIP but input from the community is welcomed.
- All methods and choices must be thoroughly documented by all groups (see Section 10 for information required).

1.2 Elements that will be defined by each modeling group (flexible)

While the core structure of the CalLMIP experiments will be standardized (see Section 1.1), modeling groups will have flexibility in defining certain methodological aspects:

- Model version: each group will select which version of their model they want to submit to CalLMIP. This includes choices such as whether LAI is prescribed or treated as a prognostic variable, and how to translate vegetation and soil characteristic information from each site to fit their model. If a group runs a variant of a land surface model that contributed to CMIP6, and can run a CMIP6-like configuration, we ask that this version be prioritized.

- Parameter selection: each group will decide which and how many parameters it wants to optimize. Parameters may vary spatially (e.g., by PFT) but should be time-invariant. The modeling groups can choose their method for parameter selection (e.g., expert elicitation, sensitivity analysis). Furthermore, each group will determine what prior parameter bounds and uncertainties to use for their parameter estimation experiments.

- Calibration variables: each group will decide which of the following variables are used in their calibration: NEE (GPP, Reco), ET/LE, and H. For groups calibrating against more than one variable, the method for multi-datastream calibration (e.g., one at a time/stepwise or all-together/simultaneous) is also left to the group's discretion.

- Parameter estimation method: each group will select its own method for calibration (e.g., gradient-based optimization, MCMC approaches, or machine learning techniques).

- Cross/multi-site parameter approach: parameter optimization must result in a set of operational parameters (e.g., for running the model globally), rather than site-specific parameter sets, although PFT-specific and soil-specific parameters, etc, are allowed. The specific method for this implementation is left to each group's discretion (e.g., hierarchical Bayesian models, individual site calibrations followed by averaging, multi-site optimizations for each PFT etc).

2. Site selection

We propose selecting 42 sites from the wider PLUMBER2 archive. These have been chosen by Gab Abramowitz and colleagues for their own model development pipeline. The main criteria for selecting these sites were: (a) length of record (meaning model spin-up issues were less likely to interfere with evaluation), and (b) data quality (less gap-filling of key driving variables - precipitation, temp, radiation, humidity, wind) (G. Abramowitz, personal communication, March 26th, 2025).

The sites are:

AU-Tum (EBF cool, 16 yrs); AU-How (WSA tropical, 15 yrs); AU-Cum (EBF, 5 yrs); AU-ASM (ENF, hot, 7 yrs); AU-GWW (SAV, 5 yrs); AU-Ctr (EBF tropical, 8 yrs); AU-Stp (GRA, 8 yrs); BR-Sa3 (EBF tropical, 3 yrs); CA-Qfo (ENF cold/snow, 7 yrs); CH-Dav (ENF, 18 yrs); CN-Din (EBF, 3 yrs); CN-Cha (MF, 3 yrs); DE-Geb (CRO, 14 yrs); DE-Gri (GRA, 11 yrs); DE-Hai (DBF, 13 yrs); DE-Tha (ENF, 17 yrs); DK-Sor (DBF, 18 yrs); FI-Hyy (ENF cold/snow, 19 yrs); FR-Gri (CVM, 9 yrs); FR-Pue (EBF, 15 yrs); GF-Guy (EBF tropical, 11 yrs); IT-Lav (ENF, 10 yrs); IT-MBo (GRA, 10 yrs); IT-Noe (CSH, 11 yrs); NL-Loo (ENF, 17 yrs); RU-Fyo (ENF, 12 yrs); US-Blo (ENF, 7 yrs); US-GLE (ENF - cold/snow, 6 yrs); US-Ha1 (MF, 15 yrs); US-Me2 (ENF, 13 yrs); US-MMS (DBF, 16 yrs); US-Myb (WET, 4 yrs); US-NR1 (ENF, 16 yrs); US-PFa (MF, 20 yrs); US-FPe (GRA, 7 yrs); US-SRM (WSA, 11 yrs); US-SRG (GRA, 6 yrs); US-Ton (WSA, 14 yrs); US-UMB (DBF, 15 yrs); US-Var (GRA, 14 yrs); US-Whs (OSH hot, 7 yrs); US-Wkg (GRA, 10 yrs).

Breakdown of IGBP vegetation types:
ENF: 11, EBF: 7, GRA: 7, DBF: 4, WSA: 3, MF: 3, SAV: 1, CRO: 1, CVM: 1, CSH: 1, WET: 1, OSH: 1 (see here for IGBP classification: https://fluxnet.org/data/badm-data-templates/igbp-classification/).

Breakdown of countries:
Australia (7), Brazil (1), Canada (1), Switzerland (1), China (2), Germany (4), Denmark (1), Finland (1), France (2), French Guiana (1), Italy (3), Netherlands (1), Russia (1), USA (16)

3. Forcing Datasets

Standardized in situ flux tower meteorological forcing data will be used to ensure consistency across sites. The forcing datasets will include SWdown, Tair, Qair, Rainf, Wind and PSurf for a number of flux tower sites.

The sites used in the first phase of CalLMIP will be based on the PLUMBER2 dataset. PLUMBER2 provides pre-selected sites from FLUXNET2015 (CC-BY-4.0 licensed sites), FLUXNET La Thuile Free-Fair-Use subset and OzFlux data, that meet quality control criteria and have minimal data-sharing restrictions. PLUMBER2 quality control steps ensure that all sites have complete metadata, including canopy height and IGBP vegetation type, and contain full years of data with minimal missing values for key variables. Furthermore, standardized gap-filling methods (e.g., LWdown from Abramowitz et al. (2012), PSurf from elevation and temperature) and FLUXNET2015-aligned energy balance corrections have been applied to these sites as part of the PLUMBER2 processing steps.

From the 170 PLUMBER2 sites, a subset (~40 sites) will be chosen by the CalLMIP planning committee for use in CalLMIP experiments. These will be chosen to maximize diversity of PFTs worldwide and ensure at least X full years of data.

The same forcing files in NetCDF format will be provided to all groups for all selected sites. Individual modeling groups will then need to modify the files to fit their model (for example, changing the variable names, adjusting or renaming the units and other attribute information, separating snow and rain etc).

4. Optimization Datasets

Optimization data will include standard *in situ* flux tower data for optimization: NEE (GPP, Reco), ET/LE, H. Datasets will be provided at daily frequency for the selected sites for the calibration period only.

Still to be discussed/decided in the 3rd workshop:

- Should we provide observation uncertainties, let groups decide how to specify those, or provide guidelines for how to specify observation uncertainties?
- Should we gap-fill the flux observations?

The same observation files in NetCDF format will be provided to all groups for all selected sites. Individual modeling groups will then need to modify the files to work with their calibration system (for example, changing the variable names, adjusting or renaming the units and other attribute information, separating snow and rain etc).

5. Required Simulations

Each group will conduct two simulations per site to ensure a consistent evaluation of model performance before and after calibration. For each of the optimized variables targeted by the group, we ask for a:

1. **prior simulation**, representing an out-of-the-box simulation using default model parameters at each site. This will serve as a baseline to assess the impact of calibration.
2. **posterior simulation**, using the optimized parameters and reflecting the improvements achieved through the calibration process.

All groups must follow the spin-up and transient protocol prior to performing the parameter calibrations (see Section 7).

Optional: Groups can also submit prior and posterior simulations at each site for additional variables not targeted in the optimization – which variables to be decided.

Optional: Groups can submit prior and posterior simulations for different model versions.

Tiered Experiments

6. Output specifications

Model results must be output at daily frequency and stored in NetCDF files, using standard CMIP variable names, and CF-compliant metadata, ensuring consistency in variable naming, units, and documentation (a table will be provided by the CalLMIP planning committee – but see this site: https://modelevaluation.org/variableStandards).

All model groups must provide prior and posterior uncertainties on both parameters and optimized variables in the NetCDF output files (to be decided).

A standardized naming system for output file names will be required (to be provided by the CalLMIP planning committee).

Each group must also provide detailed metadata alongside their results. This includes the model version, the number of parameters used in calibration, the chosen calibration method, and a summary of the calibration setup (see Section 10).

7. Spin-up and transient protocol for the parameter optimizations

Each modeling group must ensure that a full spin-up is performed that is appropriate for the variables they are optimizing. We will follow the PLUMBER2 spin-up protocol. Models will cycle over the flux tower meteorological forcing files until equilibrium has been reached in the model states of interest. For models focusing on carbon fluxes, a full spin-up to equilibrium carbon stocks (and N and P) will be required (likely 100s to 1000s of years depending on model structure). For models that are only focused on hydrological or energy balance variables and have prescribed LAI, a shorter spin-up will be sufficient (likely at least 10 years of simulation).

For carbon stocks we will spin up to equilibrium representing the year 1850 using atmospheric CO_2 concentration and N deposition levels of 285 ppm and 0.79 kg N ha-1 yr-1, respectively. The transient simulation will cover the year 1851 to the first year of the forcing data. Climate forcing during this transient simulation will continue to cycle over the available flux tower site meteorological forcing files, but models should use historical changes in atmospheric CO_2 concentration and N deposition (files will be provided).

Note 1: If groups can and want to initialize their model runs by prescribing initial conditions and states they can do that.

Note 2: the spin-up and transient protocols will be different if we do global post-optimization runs and/or site-based future simulations. See Section 9.

8. Optimization validation at site-scale

Site-level validation will be conducted for the optimized variables using data years and sites that were not included in the calibration process (exact number or percentage of years for validation to be decided by the CalLMIP planning committee). At a minimum, validation simulations using the must be performed for the same variables used in optimization. Modeling groups may choose to provide validation simulations for variables not included in the optimization.

Note 1: Additional evaluation of variables not provided for optimization will depend on data availability at the selected sites. If CalLMIP participants know of in situ data on model states (e.g., soil moisture, land surface temperature aboveground biomass or soil C stocks) that is readily available and gap-filled please let us know.

9. Additional runs/analyses

Two other potential sets of runs (with additional analyses) have been proposed for the CalLMIP 1st stage. An outline of what these additional runs would require and the protocol for doing so is provided below. During the 3rd session of the workshop we will collectively discuss whether to do either or both of these additional runs as well as the specific details.

Post-optimization global runs over the last two decades

Global scale runs (for both prior default parameters and posterior optimized parameters) would enable us to evaluate large-scale model performance. Similarly to the site-level experiments, the input data for the global simulations, such as climate forcing, atmospheric CO2, land use/cover change, N deposition etc will be the same across models and will follow the TRENDY protocol – i.e. we will use the TRENDY input data and the same spin-up and transient protocol unless the full TRENDY spin-up is not required). These input data would also be prepared by the MacBean lab and distributed by the CalLMIP planning committee. The outputs will be standardized across models including the variables required (likely will include model states and not just fluxes), time period (likely at least the last 20 years), frequency of output (likely monthly and/or annual), and the spatial resolution (0.5 degrees, 1 degree or 2 degrees). These will be decided upon by CalLMIP participants during the third session to enable as many groups to participate as possible.

The datasets and evaluation/benchmarking tools that will be used for evaluation/benchmarking of global scale simulations will be decided at a later date. Options include satellite based products and machine learning based upscaled products. Options for benchmarking and analysis include iLAMB, Ecological Forecasting Initiative (EFI) computational infrastructure, ESMValTool.

Post-optimization site-based future simulations up to 2100

Future simulations for both prior and posterior parameter sets would enable us to have an assessment of how parameter optimization is affecting our future predictions, albeit not at global scale. For these runs we would provide offline climate forcing from the recent CMIP6 ESM outputs (to be decided) from the pre-industrial to 2100, allowing a full transient run (with changing climate, CO_2, LUC) to be performed. Land use change, future CO_2 concentrations and N deposition data would be provided (to be decided).

Note 1: We note that both sets of additional simulations require different forcing data to what will be used for the optimizations. We therefore will need to be careful not to discuss model "improvement" per se as biases/uncertainties due to parameters calibrated to different forcing may propagate in unknown ways to global scale or future simulations. What we can assess is the difference between prior and posterior and how that compares to benchmark data (for global runs) or how that changes future predictions.

Note 2: We caution that global runs and new future simulations when model states have not been optimized is problematic and needs to be addressed in the optimization and/or considered when performing the additional runs/analyses.

Note 3: There is potential to compare the calibrated outputs with empirical models, *similar to PLUMBER2. This will be discussed further with Gab Abramowitz.*

10. Model/experiment information table

The following information will be required from each modeling group. We will provide a link to a google spreadsheet for modelers to fill out. The following is a general list of what we will require.

1. Model and version
2. Parameter optimization method
3. Number of parameters (targeted and total, i.e. total = the number of parameters from which the targeted/optimized parameters have been selected)
4. Which parameters and short description
5. Method of parameter selection (e.g., sensitivity analysis) (also something about the selection/screening method?)
6. Parameter bounds including PFT variation, if applicable
7. Spin-up protocol used or if initial states prescribed (and if so, using which data/method)
8. Cross/Multi-site parameter calibration approach used (i.e., calibration protocol resulting in generic enough parameters that are applicable across multiple locations/for global simulations and not limited or specific to a single site – see Section 1.2 “Cross/multi-site parameter calibration approach”)
9. Calibration variables
10. Calibration datasets
11. …

11. Additional Information

11.1 Participants

Model	Who	Institution	Email	Which variables do you plan to optimize?
CABLE-POP	Matthias Cuntz	INRAE, France	matthias.cuntz@inrae.fr	NEE, LE
ORCHIDEE	S. Beylat V. Tartaglione V. Bastrikov C. Bacour P. Peylin L. Olivera	LSCE CEA/IPSL, France	orchidas-callmip@lsce.ipsl.fr	

11.2 Potential Additional Experiments

Additional experiments	Participant(s)
Comparing parameter optimization with site forcing versus climate reanalyses	??
Compare the same optimisation method across different models	ORCHIDEE (2-3 different versions of the model)
Compare different optimization methods across the same model	CLM, ORCHIDEE (to be determined for ORCHIDEE depending on the differences obtained)
Include state variables in optimization	CLM

(e.g., soil C, soilm)	

12. Open Data Policy

All outputs from CalLMIP parameter calibration experiments will be deposited in a publicly available repository (such as …). These model outputs will be freely available for public use with an authorship and acknowledgement policy. Users of these model outputs will be required to reach out to the individuals and modeling groups who performed the experiments to offer co-authorship within X years of the model outputs being deposited in the repository. Users will also be required to cite the CalLMIP 1st Phase paper (citation will be provided).

13. CalLMIP co-authorship policy

Any individual who contributes to performing parameter calibration experiments, and/or who contributes in an intellectual capacity to the design of the experiments and analyses, and/or who advises individuals carrying out parameter calibration experiments, and/or who documents the experiment information will be included as a co-author on the paper resulting from the CalLMIP 1st Phase experiments. Co-authors will also be expected to provide feedback on the structure and framing of the paper as well as on draft(s) of the paper.

14. Future vision for CalLMIP

To be updated based on the 3rd session…