Calibrated Land Model Intercomparison Project (CalLMIP)

Phase 1 Protocol v1.0

Contact: Natasha MacBean (nmacbean@uwo.ca) and Raj Deepak S.N. (rsurulin@uwo.ca)

Background

CalLMIP is a new 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. CalLMIP is 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. Please see the CalLMIP website (https://callmip-org.github.io) “About” page for who is a member of the CalLMIP steering committee.

This initiative is targeted at global-scale land models – i.e., prognostic, process-based 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 individual model uncertainty and inter-model spread on key optimised and non-optimised variables (e.g., carbon and water stores and fluxes, aboveground biomass, soil carbon) as a result of parameter calibration.
- 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.

In Phase 1 we are taking a *‘come as you are’* approach, which is to leverage existing/current model parameter estimation systems, rather than requiring development of new systems or tests of different calibration configurations. With this approach, we hope to maintain flexibility to include a wide array of models with their existing calibration systems.

In Phase 1 we will also have a focus on **site-level calibration** focusing only on the most common observations measured at flux tower sites: net ecosystem exchange (**NEE**), latent heat flux (**Qle**) and sensible heat flux (**Qh**). Other key ecosystem variables – such as LAI, above and belowground carbon stocks and soil moisture – will be the target of future phases of CalLMIP. Regional to global scale (as well as coupled) parameter calibration experiments will also be a target for future CalLMIP Phases.

Our hope is that Phase 1 is a starting point from which additional analysis and experiments will develop. We will outline the vision for future CalLMIP phases from the CalLMIP Steering Committee and discussion at the 1st Phase Planning workshops held in March and April 2025 in a perspective article to be submitted in spring 2026.

Key Information

- CalLMIP website: https://callmip-org.github.io
    - Brief overview of everything CalLMIP, including links to Phase 1 protocol, data, and event information.
- CalLMIP GitHub site: https://github.com/callmip-org/Phase1
    - Phase 1 repository containing protocols and data (so far). Analyses will be shared via this repository as well as on the CalLMIP workspace on modelevaluation.org.
- CalLMIP workspace on modelevaluation.org
    - Information on experiments, sites included in each experiment, forcing data and participating models. Model outputs will be uploaded here.
- AIMES Land DA Working Group website: https://aimesproject.org/cal-lmip/
    - For recordings from the 1st Phase Planning and Development Workshop held in March and April 2025.

Proposed timeline for CalLMIP Phase 1

- 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.
- November 2025: Deliver inputs and finalised protocol to modeling groups. Draft of protocol paper circulated (if we decide to go ahead with this).
- November 2025: Webinar to present the first stage protocol and gather feedback on future CalLMIP phases and goals to the wider community.
- **November 2025 to early January 2026: Read through Phase 1a protocol and decide whether you would like to participate. If so, please fill out the Participant Form by January 9th 2026** (available on the CalLMIP website (https://callmip-org.github.io) under the 'Protocols and Data' → Phase 1 → Participants.
- **November to February 2026: Phase 1a "test" calibration at 1 site – model prior and posterior simulations post calibration uploaded to modelevaluation.org by February 27th 2026.**
    - *Note – during this period we will be redoing the site selection process based on the FLUXNET 2025 release in December 2025 (see Section 2.1).*
- March to July 2026: Full multi-site Phase 1 parameter calibration experiments carried out by individual modeling groups – model prior and posterior simulations uploaded to modelevaluation.org by July 31st 2026.
- August to October 2026:
    - Global scale simulations with calibrated parameters. Calibrated model simulations uploaded to modelevaluation.org by October 30th 2026.
    - Initial analysis and draft paper for site-level calibration intercomparison in discussion/collaboration with all participating model groups.
- September 2026: Presentation of initial site level calibration results at the 2nd Land Surface Modeling Summit.
- December 2026: Presentation of initial results from global post-optimization simulations at AGU 2026.
- January to April 2027:
    - Finalise site level intercomparison paper and submit.
    - Draft paper for global post-optimization intercomparison
    - Planning for possible Phase 2.

Protocol Plan

1. Overview of CalLMIP Phase 1

In this section we outline some of the components of the experiments that will be “fixed” or constant across all participating model groups followed by components that are “flexible” and can be decided by each group. We note that in an ideal world many more components would be constant across model groups; however, as noted in the Background section, in Phase 1 we are taking a *‘come as you are’* approach that we hope will enable as many model groups as possible to participate. In future phases of CalLMIP we hope to test differences between calibration systems, forcing data, parameters selected, data types included, etc. We will outline the vision for future CalLMIP phases from the CalLMIP Steering Committee and discussion at the 1st Phase Planning workshops held in March and April 2025 in a perspective article to be submitted in spring 2026.

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, including:

- No. of sites (see Section 2).
- Required calibration experiments – including an initial “Phase1a” test calibration at 1 site (see Section 3).
- Flux tower data: forcing datasets and flux observations (variable type and their associated measurement errors) – including common record length across all groups for calibration and validation (see Sections 4 and 5).
- The required output variables will be standardized across all models (see Section 6).
- Posterior uncertainties on parameters and variables optimized will be required from all models (see Section 6).
- Output file format will be standardized across models (see Section 7).
- Spin-up protocol will be partly standardized across all groups (see Section 8) for further details).
- Calibration/validation procedure (both spatial and temporal validation) will be standardized (see Section 9).

- Post optimization global simulations will be required by all groups after the full multi-site calibration (see Section 10).
- All methods and choices must be thoroughly documented by all participating groups in the Phase 1 protocol draft article (see Section 12 for information required).

Further information on all required elements is provided in the following Sections.

1.2 Elements that will be defined by each modeling group

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. These “flexible” elements of the calibration experiments are detailed in the bullet points below in this section and not discussed further in the protocol:

- 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 PFT type/fractional cover and soil texture 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, and/or if they have the version for CMIP7-FT, we ask that one or both of these versions be prioritized.
- Parameter selection: each group will decide which and how many parameters they want to optimize and which process parameters they want to include in the calibration. Parameters may vary spatially (e.g., by PFT) but should be time-invariant. Participating modeling groups can also 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.
- Parameter calibration 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, parameter ensembles etc).
- The method for including multiple data types in the calibration experiment (e.g., one at a time/stepwise or all-together/simultaneous) is left to the group’s discretion.

- Number of PFTs to be calibrated for each site: each group will decide whether they calibrate the parameters for all PFTs at a given site or only for the most dominant PFT, depending on model configuration and parameter calibration system.
- Model errors are specific to each model and need to be estimated by each group and combined with measurement errors provided, as appropriate for the parameter calibration method.
- Optional calibration experiments: groups can opt into running the additional, optional Calibration Experiments 2 to 6 (see Section 3.1).
- Certain elements of the spin-up are left to each participating model group to decide (see Section 8).

NOTE: the decisions modeling groups make for each of these points must be documented in the Phase 1 protocol draft article, which will be sent to all modeling groups that sign-up to participate (see Section 12 for further information).

2. Site selection

2.1 Overview

We have selected 22 sites from the 42 sites originally selected from the PLUMBER2 list of sites. Below is a description of the site selection process and the sites selected.

However, **PLEASE NOTE that we are planning to re-do the site selection process to add more sites/expand global coverage once FLUXNET 2025 is available (AGU, December 2025)** as our goal is maximize diversity of PFTs worldwide and ensure at least 3 full years of data. We will follow the same site selection criteria as outlined below, but we are expecting a much wider range of sites will be available: a) because of increased record lengths; and b) because all sites will have uncertainties through the harmonized flux data processing that all flux networks are working on for FLUXNET 2025. Therefore, for now (November 24th 2025) we are proceeding with Phase1a, which is a test calibration at 1 site selected from the 22 PLUMBER2 sites selected below. We will then update the sites selected in January/February 2026 before finalising the list of sites for the full Phase 1b multisite experiment, which will start in March 2026.

For each site, the following will be provided: i) gap-filled *in situ* meteorological forcing data (variables typically used in land surface models – see Section 4) from the flux tower for the entire calibration and validation period; ii) measured (i.e., “raw” or “uncorrected”) daily observations of NEE, Qle, and Qh, with measurement errors, for calibration only (see Section 5); iii) PFT type and % cover; iv) soil layer depth and

texture information; v) atmospheric CO_2 forcing data for the observation period.

2.2 Site Selection Procedure

The sites used in the first phase of CalLMIP will be based on the PLUMBER2 dataset (Abramowitz et al., 2024). 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; however, we will use only measured energy fluxes for model calibration within CalLMIP (see Section 5).

From the original 170 PLUMBER2 sites, an initial 42 sites from PLUMBER2 were 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).

We added the following additional criteria for site selection within CalLMIP based on feedback from the CalLMIP 1st Phase planning workshops held in March and April 2025 and decisions made by the CalLMIP steering committee:

1. Four year minimum record length. Within these four years there had to be at least three years worth of daily data after taking into account missing and gap-filled data with no large gaps in coverage. Only gap-filled data from NEE, Qle, and Qh was assessed to determine if there was an appropriate amount of data.
2. Netcdf files had to include uncertainties for NEE, Qle, and Qh. In the PLUMBER2 archive, only netcdf files using FLUXNET 2015 contained uncertainties. Australian sites processed using OzFlux provided no uncertainties, therefore these sites were excluded for now. Some older sites were from the LaThuille dataset, which only had uncertainties for NEE and were therefore excluded. *However, note that once we consider FLUXNET 2025 we will have access to uncertainties for all possible sites, so this criteria will be met.*
3. Some sites were further excluded because Qg (ground heat flux) was not measured. Though Qg is not a mandatory measurement within the Fluxnet processing pipelines Qg is essential for calculations of energy balance closure. EBC allowed for the calculation of an energy balance correction factor that was then used in an equation to calculate the joint uncertainty.

4. PFT and soil texture information had to be available from at least one source (see Section 2.3).

Of the 42 pre-selected PLUMBER2 sites, the following 22 sites met our selection criteria (DK-Sor will be the test site for Phase 1a):

1. CA-Qfo (Canada; 7 years; Boreal ENF)
2. CH-Dav (Switzerland; 18 years; Boreal? ENF)
3. DE-Geb (Germany; 14 years; crop)
4. DE-Gri (Germany; 11 years; NC3)
5. DE-Hai (Germany; 13 years; Boreal? DBF)
6. DE-Tha (Germany; 17 years; Boreal? ENF)
7. ***DK-Sor (Denmark; 18 years; Boreal? DBF)**
8. FI-Hyy (Finland; 19 years; Boreal ENF)
9. FR-Pue (France; 15 years; Temperate EBF)
10. IT-Lav (Italy; 10 years; Boreal? ENF)
11. IT-MBo (Italy; 10 years; NC3)
12. IT-Noe (Italy; 11 years; Temperate EBF/SH)
13. NL-Loo (Netherlands; 17 years; Temperate ENF)
14. RU-Fyo (Russia; 12 years; Boreal ENF)
15. US-MMS (US; 16 years; Temperate DBF)
16. US-NR1 (US; 16 years; Boreal? ENF)
17. US-SRM (US; 11 years; Mixed shrub-grass – Temperate DBF and NC3/NC4)
18. US-SRG (US; 6 years; NC3/NC4)
19. US-Ton (US; 14 years; Mixed shrub- grass – Temperate DBF and NC3)
20. US-Var (US; 14 years; NC3)
21. US-Whs (US; 7 years; Mixed shrub-grass – Temperate EBF and NC3/NC4)
22. US-Wkg (US; 10; NC3/NC4)

Summary of PFTs:

- 8 evergreen needleleaf forests (3 boreal, 4 probably boreal, 1 temperate)
- 3 deciduous broadleaf forest sites (1 temperate and 2 probably boreal)
- 2 evergreen broadleaf tree/shrub sites (temperate; Mediterranean)
- 3 mixed shrub-grass (SAV) sites (2 temperate deciduous broadleaf, 1 temperate evergreen broadleaf; 1 C3, 2 C3/C4)
- 3 C3 grass sites
- 2 mixed C3/C4 grass sites
- 1 crop site

Figure 1. Location of the 22 selected sites from the original 42 selected by Gab Abramowitz

The coverage of these sites is shown in Figure 1. However, NOTE that as detailed in Section 2.1, **we will expand the site selection once FLUXNET 2025 is available and has been processed for LSM use within modelevaluation.org by Gab Abramovitz. Once that dataset is available we will update the site selection for the full multi-site calibration in Phase 1b, which is planned to begin in March 2026**. This will provide us with expanded PFT and global coverage that we seek for the Phase 1 calibration experiments.

2.3 PFT and soil texture information

PFT information was not always available as % fractional cover of land surface model PFTs within the PLUMBER2 netcdf files; therefore, we consulted external sources where needed. External sources included the Shi et al., (2024) dataset (Shi et al., 2025), information from flux tower site websites and publications, and collaboration with Vincent Tartaglione and Cédric Bacour at LSCE who have performed their own analysis of the PLUMBER2 PFT information (more details will be provided in the Phase 1 protocol draft paper). To ensure that a given site had correct PFT information we confirmed that two out of three sources were consistent in their PFT information; otherwise, additional information or discussing with the PIs is needed.

Information on soil composition was collected from netcdf's when possible but otherwise sourced from Shi et al. (2024).

PFT type and % cover, soil layer depth and soil texture information are provided in the global attributes of the flux observation file.

3. Calibration Experiments

3.1 Variables included

Daily NEE, Qle and Qh data will be provided for all sites with their associated uncertainties. The following is a list of potential calibration experiments with these data; however, **only the 1st calibration experiment is required** for model groups participating in CalLMIP Phase 1.

1. **NEE, LE (Qle) and H (Qh) (required)**
2. NEE only (optional)
3. LE only (optional)
4. H only (optional)
5. NEE and LE (optional
6. LE and H (optional)

Note: while all participating groups must run Calibration Experiment 1, the method for combining all 3 datastreams into the same calibration experiment (e.g., one at a time/stepwise or all-together/simultaneous) is left to each group's discretion.

3.2 Test calibration at 1 site (Phase 1a)

Phase 1a is designed to help each model group get the experimental set-up established for the Phase 1 experiments. For this we ask you to perform the calibration following the full protocol for just one site (DK-Sor) and upload your simulations to modelevaluation.org (see Section 11). This will also allow us to verify that modelevaluation.org processing script can read and process your model outputs.

3.3 Full multi-site calibration (Phase 1b)

Once all model groups have tested their CalLMIP set-up in Phase 1a the full multi-site calibration experiments will begin in Phase 1b. We will also update the sites to be included in Phase 1b once FLUXNET 2025 becomes available in December 2026 (see Section 2.1 for further details). In the event that FLUXNET 2025 is not available by early 2026, or that the processing required to get files ready for use with LSMs is not completed by spring 2026, we will proceed with the 22 sites selected from the PLUMBER2 files (see Section 2.2) despite the lack of global coverage. We would then aim to update the selected sites to have a wider global coverage in a future phase of CalLMIP.

4. Forcing Datasets

A common climate forcing dataset will be used by all participating model groups to ensure consistency across sites. The forcing data files were prepared for PLUMBER2 by Gab Abramovitz and will be distributed to participating modeling groups via modelevaluation.org (see Section 11). These data are the in situ flux tower meteorological measurements collected at each of the selected sites. The forcing datasets include LWdown, SWdown, Tair, Qair, Rainf, Wind and PSurf for each site, among other variables. We will also provide the NCO code needed to convert Precip into Rainf and Snowf in the Phase 1a/b Experiment within the CalLMIP workspace on modelevaluation.org for those modelers who can use/need both variables. See Section 11 for how to navigate to the Phase 1a/b experiments on modelevaluation.org.

For models that predict a prognostic leaf area index (LAI) we would prefer you to use that model capability. But for models without prognostic LAI, models should use the remotely sensed LAI time series provided for each site in the PLUMBER2 netcdf forcing files. The LAI time series were derived from MODIS (8-daily MCD15A2H product, 2002-2019) or Copernicus Global Land Service (monthly, 1999-2017), with one of these chosen for each site, at the same time step size as meteorological forcing, based on a site-by-site analysis considering plausibility and some in-situ data (Abramowitz et al., 2024; PLUMBER2 protocol). The preferred product has a variable name “LAI”, the other “LAI alternative”. Time-varying LAI is provided for the time period covered by the remotely sensed products and otherwise a climatology constructed from all available years is used (Abramowitz et al., 2025; PLUMBER2 protocol).

Atmospheric CO_2 and N deposition forcing data for the historical/transient run (pre-calibration) will also be provided via the GitHub CalLMIP site in Phase1a repository Data folder https://github.com/callmip-org/Phase1/tree/main/Data).

The same meteorological 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).

5. Flux Data for Optimization

Optimization data include standard *in situ* flux tower data for optimization: **NEE, Qle and Qh**. Datasets are provided at daily frequency for the selected sites for the calibration period only. One year of data has been retained for temporal validation; therefore, the flux data file is one year shorter than the meteorological forcing file.

The flux data file is not available on modelevaluation.org (unlike the meteorological forcing file. Instead, the flux data file can be downloaded from https://github.com/callmip-org/Phase1/tree/main/Data/Phase1a-test.

Daily measurement errors are also provided to all groups based on FLUXNET processing (Pastorello et al., 2020). The uncertainty for each variable is designated by _uc. *These measurement errors need to be combined with modelling errors for the observation error covariance matrix in certain DA methods (e.g. variational)*. Each modeling group will decide on their own method for estimating model error.

Only measured flux data (i.e. non gap-filled, no energy balance closure corrected data) are provided. Days that contain more than 25% gap-filled half-hourly data points were excluded (set to NaNs). Days were also excluded if uncertainty estimates were not available. The daily missing % is provided in the flux data netcdf file for each variable.

The equivalent FLUXNET 2015 name for both the variables and their uncertainties can be found in the local attribute information within the netcdf file. For more information on these data please see Pastorello et al. (2020).

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

NOTE: PFT type and % cover, soil layer depth and soil texture information are provided in the global attributes of the flux observation file.

6. Required Outputs

- Each group will conduct two simulations per site for each experiment they perform (note: 1 experiment is required – see Section 3.1) to ensure a consistent evaluation of model performance before and after calibration:
    i. **prior simulation**, representing an out-of-the-box simulation using default model parameters in the model version being used at each site (i.e. no prior parameter tuning or updating of the model to fit the data). This will serve as a baseline to assess the impact of calibration.
    ii. **posterior simulation**, using the optimized parameters and reflecting the improvements achieved through the calibration process.

- All groups must output the following variables (per grid cell, not per PFT, age cohort or otherwise) with variable names according to the output file specification in Section 7:
    - NEE
    - Qle
    - Qh
    - GPP
    - Total Ecosystem Respiration (Reco)
    - Transpiration

    - Bare Soil Evaporation
    - Ground Heat Flux
    - Land Surface Temperature
    - Total Column Soil Moisture
    - LAI
    - Total Aboveground Biomass
    - Total Soil Carbon

- All groups must provide posterior uncertainties (or ensemble simulations) on variables optimised within the model output file.
- Model groups do not need to provide posterior parameter values or their uncertainties. However, we ask that modeling groups retain this data in case we (as a community) decide that this information could be useful for further analysis or discussion.
- All groups must output files according to output file specifications listed in Section 7.
- All groups must follow the spin-up and transient protocol prior to performing the parameter calibrations (see Section 8).
- Each group must also provide detailed metadata alongside their results. This information will be input directly into a draft paper documenting the 1st phase protocol, which will be sent to all participating groups when they sign up. Information required is listed in Section 12 and in the relevant methods and supplementary sections of the draft protocol paper.

Optional:

- Groups using ensemble methods can submit prior and posterior ensemble simulations; however, this is not required for groups not using ensemble methods. The choice of how many ensembles to include, and the method for selecting those ensembles, is left to each model group.
- Groups may perform the experiments with different model versions. In that case, we ask model groups to submit all required outputs for each model version and follow the file naming specifications listed in Section 7.

7. Output file specifications

Model results must be output at daily frequency and stored in NetCDF files, using standard ALMA variable naming convention, and CF-compliant metadata, ensuring

consistency in variable naming, units, and documentation. A list of standard names is provided here: https://modelevaluation.org/variableStandards. These files will then be uploaded to modelevaluation.org (see Section 11.5).

All output variables (optimized and additional – see Section 6) will be stored in the same output file as well as uncertainties on optimized variables.

A standardized naming system for output file names is required:
<*model_name*>.<*model_version*>_Expt<*CalLMIP_expt_no.*>_<*SiteName*>_<*Cal/Val*>_<*Prior/Posterior for calibration or Temporal/Spatial for validation*>.nc

E.g. the prior simulations for the DK-Sor site for the required experiment for CLM v5 would be: CLM.5_Expt1_DK-Sor_Cal_Prior.nc

Files containing global simulations (see Section 10) will have the following naming convention:

<*model_name*>.<*model_version*>_global_<*resolution*>_<*output_frequency*>_<*YYYY-YYYY*>_<*Prior/Posterior*>.nc

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

Each modeling group must ensure that a full spin-up is performed and that models start from a stable, physically realistic state, preventing biases from arbitrary initial conditions. To keep the spin-up consistent across all models we are following the PLUMBER2 spin-up protocol for carbon (unless the group has a set functionality to prescribe initial conditions/states).

Models will cycle over the flux tower meteorological forcing files until equilibrium has been reached in the carbon stocks (and N and P included) (likely 100s to 1000s of years depending on model structure). We will spin up to equilibrium in carbon stocks 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 yearly changes in atmospheric CO_2 concentration and N deposition (see Section 4).

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

Note 2: Forest age/disturbance that may be required for example by vegetation demography/cohort models is not provided. If needed, we can discuss further with relevant modeling groups.

9. Validation at site-scale

Temporal validation will be conducted for the optimized variables using the last year for each site. Spatial validation will be conducted by excluding 10-20% sites from the calibration (depending on how many sites we select from FLUXNET 2025). Meteorological forcing files will be provided for the temporal validation time period and spatial validation additional sites, but no flux data will be provided. Therefore the forcing file will contain both calibration and validation forcing data together in the same file, but the flux data file will contain one year less data. Model outputs and output file specifications must be the same for the validation runs as for the prior and posterior simulations (see Sections 6 and 7).

Note 1: Additional evaluation of variables not provided for optimization (as opposed to just an assessment of reduction in inter-model spread) 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 are readily available, please let us know.

10. Post-calibration global simulations

Global scale runs (for both prior default parameters and posterior optimized parameters) will be performed after the Phase 1b full multi-site parameter calibration experiment. These global scale simulations will allow us to evaluate model performance over the larger scale regions/continents that we typically look at in climate change impact studies. Similarly to the site-level experiments, the input data for the global simulations, such as climate forcing, atmospheric CO_2, 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). Links to these input data will be provided on the CalLMIP website at the start of the global simulations (see Proposed Timeline).

The outputs will be standardized across models including the variables required (see Section 6), 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 during a virtual meeting with CalLMIP participant model groups during the summer 2026.

The model outputs will follow the same variable and file format naming conventions as specified in Section 7 and will be uploaded to modelevaluation.org and evaluated using the iLAMB model benchmarking package which is integrated into modelevaluation.org.

Note 1: We note that these 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. If participating model groups are interested, we may run an additional set of site level calibrations with the TRENDY forcing data to assess what impact the different forcing has on parameter calibration performance.

Note 2: We caution that global runs 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.

11. Getting set-up on modelevaluation.org

11.1 Creating an account and accessing the CalLMIP workspace

You will need an account to be able to access the CalLMIP workspace. If you haven't already, please create an account directly on the website.

- The CalLMIP workspace is public and therefore should be automatically shared with you once you sign up to modelevaluation.org.
- To find the CalLMIP workspace find the button in darker blue to the right of tabs at the top of the page titled 'Current Workspace'. Once you click the button you will be taken to a new page where there are workspaces that have been shared with you.
- Scroll down to find CalLMIP.workspace to Click the CalLMIP workspace and you should see that darker blue tab at the top of the page change to say 'Current Workspace: CalLMIP'. You should now be able view the data, sites, experiments and model output options available within that workspace.

11.2 Finding CalLMIP experiment information

Information on all experiments being performed within CalLMIP Phase 1 can/will be found within the CalLMIP workspace on modelevaluation.org (see Section 11.1 to create an account on modelevaluation.org and find the CalLMIP workspace).

- To access the information on the Phase 1a/b experiment click on the 'Experiments' tab on the top of the page and go to 'In Current Workspace' on the drop down menu. This will take you to a list of all the experiments for CalLMIP. Currently only the Phase 1a experiment is available to view (November 24th 2025). This will be updated once the Phase 1b experiments are finalised by the beginning of March 2026 (see Proposed Timeline).

- Click on the CalLMIP Phase 1a experiment from the list to see general information about the experiment.
- Additional information – such as how to partition Precip into Rainf and Snowf – is also provided on this experiment page.

11.3 Finding site and data information

Information on all sites and forcing files used within the CalLMIP project can be found within the CalLMIP workspace on modelevaluation.org (see Section 11.1 to create an account on modelevaluation.org and find the CalLMIP workspace).

- To access site information and meteorological forcing datasets available for each site click on the 'Datasets' tab on the top of the page and go to 'In Current Workspace' on the drop down menu. This will take you to a list with information and met data on all the sites and forcing datasets within the CalLMIP experiments
- Note 1: there is only one site available for the Phase1a test calibration – DK-Sor. The number of sites available will be updated once we move on to the full multi-site calibration in Phase1b).
- Note 2: Flux observation datasets are NOT available on modelevaluation.org. The flux data will be shared via the CalLMIP GitHub site. See Section 5 for further details.

11.4 Creating a Model Profile

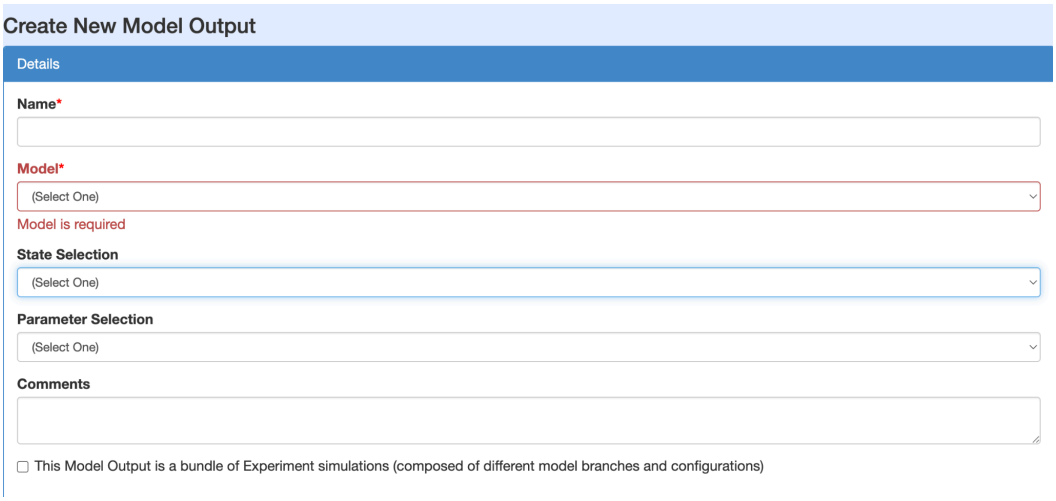
All model simulations (prior and posterior, calibration and validation) generated will be uploaded to modelevaluation.org. Every participating model group will need to create a new model profile.

- To create a new model profile, click on the 'Model Profiles Tab' and then select 'Create New Model Profile' from the drop down.
- Input Model name (and version if you are only submitting simulations from one model version)
- Select whether your model profile is visible publicly or only within the CalLMIP workspace.
- Provide a url(s) for the model.
- Provide relevant references for both the model and calibration system. However, please note that we will collect most of the information about the model and calibration within the Phase 1 protocol draft paper. See Section 12 for details that will be required.
- Provide any additional comments that you think may be helpful, but see note in previous bullet point about where we collect most of the information relating to the model and calibration system.
- Click save and now this model profile can be repeatedly accessed within the 'Model Outputs' tab on the top of the page. You can upload as many model outputs (including prior and posterior, calibrated and validated, required and optional experiments, different versions) under this created model profile, as needed.
- Note: all participating model profiles are available to view under 'Model Profiles' → 'In Current Workspace'.

11.5 Uploading Model Outputs

Once you have created a model profile (see Section 11.4) and have completed the experiments required for Phase 1a/b, you can then upload your model outputs. NOTE: you can upload all the files for each Phase (1a and 1b) under one new model output *for each site*.

- To upload your model output for each site, click on the 'Model Outputs' tab at the top and select the option to 'Upload Model Output'. *Note you will need to do this for each site (for Phase 1a this will just be for the one test site).*
- You should now see this page (top 5 information boxes shown)



- Within the 'Model*' dropdown you should be able to see the model you have previously created from the 'Create New Model Profiles'.
- Under 'State selection' select the relevant description. If your method is not listed please select 'other' and ensure this is explicitly detailed within the comments section at the bottom of the page.

- Under 'Parameter Selection' select 'automated calibration' given at least your posterior simulations are with calibrated parameters.

- Upload all required model outputs (see Section 6 and 7) and any ancillary files as necessary.
- **Ensure you 'Save' this output**.

12. Model/experiment information table

The following information will be required from each modeling group. This information will be input directly into a CalLMIP 1st Phase protocol draft paper that we aim to publish in GMD (see Proposed Timeline). This draft paper will be shared with all participating model groups once they have registered on the CalLMIP website. The information required includes (but please see the draft paper for the most up-to-date list):

i. Brief description of the model including the model version(s) plus all relevant references.
ii. What parameter estimation method are you using? Detailed steps and equations should be documented in your models' section below. Any other details on configuration/set-up should also be included (e.g., no. of ensembles etc).
iii. Description of parameter selection method (including SA if SA was performed) and any additional screening steps that were applied. Include how many parameters were selected out of a total X number of possible model parameters to be calibrated.
iv. Table of parameters included in the calibration, including their short (model) and long name, description, process associated with the parameter, prior/default value, prior uncertainties, ranked sensitivity (if an SA was performed).
v. Method for the spin-up beyond what is outlined above?
vi. Method for accounting for potentially biased initial conditions (soil C stocks etc) - are they included in the parameter calibration or prescribed for example?
vii. Method for calibrating with multiple datastreams (experiments iv to vi) what is your method/approach for multi data stream assimilation(e.g., step-wise vs simultaneous)? I.e. How are these datasets combined in the cost/objective function?
viii. What is your method for generating an operational set of parameters for running the model globally? (e.g. multi-site optimisation per PFT or parameter ensembles?)
ix. Do you calibrate all PFTs for a given site or only the most dominant PFT (where applicable)?
x. What other sources of uncertainty are being included/propagated in the calibration (e.g., initial condition, driver, boundary, demographic stochasticity, random process error).
xi. Which optional experiments are you participating in (e.g., NEE only, LE only etc; see protocol Section 3.1).
xii. Did you prescribe LAI from the satellite time series provided in the forcing file, or did you use a prognostic LAI?

13. Open Data Policy

All outputs from CalLMIP parameter calibration experiments are hosted on modelevaluation.org. These model outputs will be freely available for public use after 1 year of publication under the Creative Commons Attribution 4.0 International (CC-BY-4) license by registering on modelevaluation.org. Use of model outputs must include an acknowledgement of the model group that generated the model outputs and citation of the relevant model specific and CalLMIP papers. The full acknowledgement and citation

policy (including model specific acknowledgements/references) will be provided on the CalLMIP website and in the public repository.

14. CalLMIP co-authorship policy

We recognise that global land surface model parameter calibration systems often require a considerable number of people – often early career researchers – for their development and indeed for carrying out the simulations/calibrations required by this voluntary CalLMIP initiative. Therefore, our primary goal is to ensure all those who contributed significantly to enabling CalLMIP to happen will be rewarded by co-authorship on CalLMIP papers.
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. We are working on a more detailed 'CRediT' contributor matrix (with 'Contributor, Roles, Taxonomy as columns and people as rows) with co-authorship requiring a certain number of those items in the coming weeks (November 24th 2026) so that we can share that with participating groups as they sign up. However, we note that we want to offer participating model groups as much autonomy and flexibility as possible when deciding who should be a co-author on the paper, while encouraging groups to be fair in deciding who co-authors should be.

References

Abramowitz, G., Ukkola, A., Hobeichi, S., Cranko Page, J., Lipson, M., De Kauwe, M. G., Green, S., Brenner, C., Frame, J., Nearing, G., Clark, M., Best, M., Anthoni, P., Arduini, G., Boussetta, S., Caldararu, S., Cho, K., Cuntz, M., Fairbairn, D., … Zeng, Y. (2024). On the predictability of turbulent fluxes from land: PLUMBER2 MIP experimental description and preliminary results. *Biogeosciences*, *21*(23), 5517–5538.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Reichstein, M., Ribeca, A., van Ingen, C., Vuichard, N., Zhang, L., Amiro, B., Ammann, C., … Papale, D. (2020). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, *7*(1), 225.

Shi, J., Yuan, H., Lin, W., Dong, W., and Dai, Y. (2024) A flux tower site attribute dataset intended for land surface modeling, Zenodo [data set], ttps://doi.org/10.5281/zenodo.12596218.

Shi, J., Yuan, H., Lin, W., Dong, W., Liang, H., Liu, Z., Zeng, J., Zhang, H., Wei, N., Wei, Z., Zhang, S., Liu, S., Lu, X., & Dai, Y. (2025). A flux tower site attribute dataset intended for land surface modeling. *Earth System Science Data*, *17*(1), 117–134.