March 9, 2015

PalEON Model-Data Inter-comparison Project
Phase 1a: Site-level comparison

Summary: This document outlines the protocol for running the PalEON site-level model inter-comparison for **Phase 1a**, which we encourage modeling teams to complete attempts with the bias-corrected met drivers and submit by **April 15th, 2015.** Phase 1b regional runs across the entire PalEON spatial domain has been postponed with no current target completion date. Phase 2 will commence this fall and will actively assimilate the PalEON data products to constrain the state variables in ecosystem models. In Phase 3, we will use site-level data to improve the model performance by using data assimilation to constrain model parameters and structure**.**

Phase 1a is a site-level model inter-comparison using the downscaled CCSM4 model and CRUNCEP as meteorological drivers from years 850-2010 at the three PalEON Highly Integrated Proxy Sites (HIPS) and three PalEON intensive pollen sites:

Site	PalEON	Lon	Lat
PHA: Harvard Forest	HIPS	-72.18	42.54
PHO: Howland Forest	HIPS	-68.73	45.25
PUN: UNDERC	HIPS	-89.53	46.22
PBL: Billy's Lake	Pollen	-94.58	46.28
PDL: Deming Lake	Pollen	-95.17	47.17
PMB: Minden Bog	Pollen	-82.83	43.61

Table 1. PalEON Phase 1a MIP sites: The PalEON Phase1a MIP sites span a wide geographic and climate gradient. Half are Highly Integrated Proxy Sites (HIPS) within the PalEON research design, and half are intensive pollen sampling sites. Note that the coordinates for sites correspond to grid cells within the regional 0.5-degree PalEON domain, not the precise location of individual sites.

The goals of Phase 1a are to:

1) Compare model spread across 6 sites with excellent paleo and modern data.
2) Analyze model performance at the three HIPS sites prior to any data assimilation.
3) Identify any potential areas of difficulty before embarking on regional runs across the entire PalEON spatial domain.

Phase 1a Model Simulations

PRIORITY #1:

1. PHA/PHO/PUN/PBL/PDL/PMB: **New site-level runs over the PalEON time period**
This set of runs will be conducted with a common set of bias-corrected met drivers, parsed together from the downscaled CCSM4 climate model output during the past1000 simulation (850-1849) and historical simulation (1850-1900), and the CRUNCEP dataset (1901-2010). Model output should be saved as monthly mean data. "Validation" data will not be available until after runs are submitted, so calibration is not possible. If any "tuning" is done please also submit the original baseline run and document what changes were made. Models should run with environmental driver files, including land-use change files, and turning fire on wherever possible.

2. SPHA/SPHO/SPUN/SPBL/SPDL/SPMB: Sub-daily snapshots
For models capable of simulating a sub-daily timestep, we ask that you submit high temporal-resolution model output (ideally 3-hourly) for the time periods 1600-1609, 1850-1859, and 1996-2005 for the same set of driver data described above.

PRIORITY #2:

3. UPHA/UPHO/UPUN/UPBL/UPDL/UPMB: Un-bias-corrected site-level runs
This set of runs will be conducted with the set of un-bias-corrected met drivers, to examine the sensitivity of the models to differences in reconstructed climate. The un-bias-corrected meteorological drivers were the same set used in preparation for the Berkeley Annual Meeting in December 2014 (phase1a_met_drivers_v4.1), but with unit errors fixed and now produced for all six sites. They represent the same time period and model simulations as Step 1. Models should run with land-use change files and turning fire on wherever possible.

4. PST: **Potential vegetation steady-state snapshot**
If there is one available, model teams can submit a steady-state "potential vegetation" snapshot of model output for the region based on running their model to equilibrium by cycling modern meteorology. These runs should represent pre-settlement conditions, and could come from existing model output, but should have a resolution no coarser than 1 degree. An example would be the MsTMIP steady-state initial conditions (1840) for the globe (0.5 deg, RG1) or North America (0.25 deg, RR1). Please submit 10 years of simulation to capture inter-annual variability.

Simulation naming convention:
Although models may differ in their native output frequency, files for Simulations 1 and 3 should be submitted as **100-year files of monthly mean data** for archiving and analysis, with **separate files for sub-daily output** from Simulation 2 for 1600-1609, 1850-1859, and 1996-2005. See Table 5 for a list of output specifications.

Please use the following output file naming convention for all PalEON simulations: **[simulation acronym].[model acronym].[earlisest year].nc** e.g. PHA.ED2.1500.nc

Simulation Name	**Paleoclimate Driver**	**Notes**
1. PHA/PHO/PUN/ PBL/PDL/PMB	Bias-corrected CCSM4 past1000 + CCSM4 historic + CRUNCEP	Time range: 850 – 2010 A.D. Dataset downscaled and bias-corrected based on CRUNCEP.
2. SPHA/SPHO/SPUN/ SPBL/SPDL/SPMB	Same as above	Sub-daily output for three decade-long snapshots: 1600-1609, 1850-1859, and 1996-2005.
3. UPHA/UPHO/UPUN/ UPBL/UPDL/UPMB	Original CCSM4 past1000 + CCSM4 historic + CRUNCEP	Time range: 850 – 2010 A.D. Dataset downscaled based on CRUNCEP but not bias-corrected.
4. PST	As available	Potential steady-state snapshot with modern climate

Table 2: List of simulations to run for Phase 1a MIP. See "Meteorological Driver Datasets" for detailed description of these drivers.

Model Spin-up & Initial Conditions

All simulations (P, SP, and UP site-level runs) should be spun up to steady-state initial conditions prior to the start of the PalEON simulations through time. In all cases models should **use the oldest 20 years of the meteorological data for the spin-up period** (e.g., 850-869). CO2 concentrations should be set at 277ppm, the mean of the 850-869 time period from the reconstructed CO2 concentration driver dataset. The use of a semi-analytical spinup (SAS) approach[1] is both permissible and encouraged. Please note that the additional variables required for the SAS and traceability framework[2] have been added to the protocol as encouraged outputs.

[1] Xia, J. Y., Y.Q. Luo, Y.-P. Wang, E. S. Weng, O. Harauk. 2012. A semi-analytical solution to accelerate spin-up of a coupled carbon and nitrogen land model to steady state. Geoscientific Model Development 5:1259-1271.

[2] Xia, J. Y., Y.Q. Luo, Y.-P. Wang, O. Harauk. 2013. Traceable components of terrestrial carbon storage capacity in biogeochemical models. Global Change Biology 19:2104-2116.

At the start of spin up, any prognostic soil temperatures should be initialized to the mean temperature of the 20 years of data that are being repeated. Canopy temperatures and canopy air space temperatures, pressures, and humidity should be initialized to the first observation in the met files. Biogeochemical pools and all other prognostic variables can be initialized as best suited for your model, however for models that possess a dynamics vegetation component, we strongly discourage the use of the Ramankutty and Foley biome map, or other potential vegetation maps, as a number of our hypotheses will involve comparison to these data sets and we want the dynamic models to be independent when possible. Otherwise, models with static vegetation will use the Ramankutty and Foley map to specify biome.

Meteorological & Environmental Driver Datasets

Downloading Instructions:

All simulations should be run using the standardized driver data, which can be found in the iPlant Discovery Environment:

https://de.iplantcollaborative.org/de/?type=data&folder=/iplant/home/crollinson/paleon

To gain access to the iPlant repository, you will need to register and then contact one of the modeling personnel listed at the end of this document so you can be given access to the PalEON group modeling repository.

Meteorological driver data can be found in the “phase1a_met_drivers” folder. **The most recent bias-corrected data as of March 2015 is phase1a_met_drivers_v4.2.**

A reminder: the bias-corrected dataset is the primary priority, and the un-bias-corrected runs are a lower priority if you did not complete them in 2014, but would still be very useful for the MIP if you have time. The un-bias-corrected meteorological drivers used in 2014 runs are listed as phase1a_met_drivers_v4 in the “OLD versions” folder). The .tar.bz2 files are 0.5 GB each and the unzipped directories of met data are ~19GB each. There is also **one set of environmental driver data (phase1a_env_drivers_v2.zip)**[3] for all simulations found in the “phase1a_env_drivers” folder on iPlant that is 1MB zipped, and 1.3MB unzipped.

[3] Updated 20 March 2014 to include a monthly CO_2 file with seasonal variation.

Meteorological data is given as a single .tar.bz2 file that can be downloaded from iPlant by checking that file and then selecting “simple download” in the “Download” menu in the discovery environment. iPlant also offers iDrop, a desktop app that can be used to download or upload data. More information about iDrop can be found at: https://pods.iplantcollaborative.org/wiki/display/DS/Using+iDrop+Desktop

Summary of meteorological drivers:

Within the site-level zip files there are folders corresponding to the 7 meteorological variables: lwdown, precipf, psurf, qair, swdown, tair, and wind (Table 3). Each variable folder contains monthly files of 6-hourly netCDF driver data in GMT, where each value represents the average of the subsequent 6 hours. To manipulate the meteorological drivers by renaming variables (for example to MsTMIP format) please see the code **convert_met_rename.sh** in the “Useful Scripts” folder inside of the “phase1a_met_drivers” folder on iPlant. The netCDF files should be self-documenting and include descriptions of the variables, units, and timesteps. A README file is also available that summarizes the changes to the met drivers since August 2014. Version 4 data (winter 2014) was not bias-corrected, whereas v4.1 was biascorrected, but with incorrect leap years. Version 4.2 is the current version as of March 2015 and is bias-corrected and has correct leap years.

ALMA Name (CF equiv.)	ALMA Definition	Units	Precision
lat (lat)	latitude	degrees_north	float
lon (lon)	longitude	degrees_east	float
time (time)	time	Calendar date (Gregorian proleptic)	double
lwdown [time,lat,lon] (fldlw)	Incident longwave radiation averaged over the time step of the forcing data	W m-2	float
swdown [time,lat,lon] (fldsw)	Incident radiation in the shortwave part of the spectrum averaged over the time step of the forcing data	W m-2	float
precipf [time,lat,lon] (rainfall_flux + snowfall_flux)	The per unit area and time precipitation representing the sum of convective rainfall, stratiform rainfall, and snowfall	kg m-2 s-1	float
psurf [time,lat,lon] (ps)	Pressure at the surface	Pa	float
qair [time,lat,lon] (q)	Specific humidity measured at the lowest level of the atmosphere	kg kg-1	float
tair [time,lat,lon] (air_temperature)	2 meter near surface air temperature	K	float
wind [time,lat,lon]	Wind speed measured with a vertical coordinate in height of 10 m	m s-1	float

Table 3. Meteorological driver dataset: All variable names follow the ALMA metadata conventions where possible (http://www.lmd.jussieu.fr/~polcher/ALMA/convention_input_3.html). We also reference the ALMA names to their CF equivalents (http://cf-pcmdi.llnl.gov/documents/cf-standard-names/standard-name-table/16/cf-standard-name-table.html).

Meteorological driver methods:

The CCSM4 meteorological driver data were concatenated from two model runs: a fully coupled simulation for the Last Millennium PMIP3 inter-comparison (past1000, 850-1849) and a fully coupled simulation for AR-5 (historical, 1850-1900). The CRUNCEP dataset picks up from 1901-2010 so that the PalEON meteorological drivers span 850-2010. The past1000 run was initialized with an 1850 control run. Total solar irradiance was derived from a calibrated reconstruction based on multiple datasets. GHG forcing (CO_2, CH_4, N_2O) was developed from Gavin Schmidt's (NASA Goddard) 1850-2000 dataset with pre-Industrial spline fit to ice cores from Law Dome Ice data. Volcanic aerosol forcing represented multiple eruptions and was derived from the Gao-Robock-Ammann dataset. The CCSM4 historical driver data were derived from a simulation part of the IPCC CMIP5 experiment. Forcings for the historical simulations are extensively documented at: http://cmip-pcmdi.llnl.gov/cmip5/forcing.html

The CCSM4 model output for both the past1000 and historical data were downscaled spatially and temporally using an artificial neural network (ANN) method outlined in Kumar et al 2010[4]. Briefly, the ANN uses the 6-hourly 0.5-degree CRUNCEP dataset to downscale the ~2-degree monthly mean time series output from the CCSM4 model. Due to the mis-match between model and data means at transitional periods, particularly the switch from the CCSM4 model to CRUNCEP data at 1901, model output was bias-corrected for temperature, radiation (lwdown and swdown), precipitation, and specific humidity. The bias was calculated by comparing the CRUNCEP dataset from 1961-1990 to the downscaled CCSM4 model output during the same time period. Temperature was corrected in the CCSM4 past1000 and historical output as an additive bias, and radiation and precipitation were corrected as a ratio bias. The model output for specific humidity had an additional abrupt jump at 1850 during the switch from the past1000 to historical CCSM4 simulations. To smooth this transition in order to avoid erroneous simulations in the PalEON MIP, we first bias-corrected the historical CCSM4 simulation to the CRUNCEP dataset from 1961-1990, and then assumed that the monthly mean for the first 15 years of the historical simulation (1850-1864) should match that of the last 15 years of the past1000 simulation (1835-1849). We then bias-corrected the past1000 CCSM4 simulation for specific humidity based with a ratio approach. Details on this bias-correction are in the document PalEON_met_biascorr posted on the PalEON wiki.

[4] Kumar, J., Brooks, B.-G.J., Thornton, P.E., & Dietze, M.C. 2012. Sub-daily statistical downscaling of meteorological variables using neural networks. Procedia Computer Science 9:887-896.

Environmental Driver Datasets

The environmental driver datasets used in the Phase 1a site level MIP are annual CO2 concentration, soil characteristics from the Harmonized World Soil Database (HWSD), biome type (ONLY for models with prescribed vegetation) from the Ramankutty and Foley database, and nitrogen deposition from the enhanced Dentener et al 2006 dataset developed for MsTMIP (Table 4). All environmental driver files are zipped in **phase1a_env_drivers.zip** on the iPlant site. Annual CO2 data are within one file **paleon_co2_mix.nc**, site-level HWSD characteristics and biome map have been extracted to **PalEON_Phase1a_sites.csv** and the site-level environmental drivers for nitrogen deposition and land-use change have been extracted to the folders **site_nitrogen/** and **site_lulc/** where the acronyms in the file names correspond to the site names in Table 1. Please see the README file in the site_lulc folder for interpretations of the land-use transition codes.

Data Type	**Source**	**Details**
CO2	PMIP3 + NOAA	Developed by splicing the PMIP-3 Law Dome composite time series (0-2000), and NOAA's Mauna Loa baseline CO_2 observatory data (2001-2010)
Land-Use	Hurtt Land-Use History A, version 1	These data are not to be confused with the Hurtt (2006) data. These were developed by Hurtt and others using the same general methodology as Hurtt (2006) but instead using the latest HYDE 3 historical data set for crop, pasture, and urban area (1500-2005).
Soil	HWSD	Harmonized World Soil Database, extracted for each site into PalEON_Phase1a_sites.csv
Biome	Rammankutty and Foley	0.5 degree gridded product, extracted for each site into PalEON_Phase1a_sites.csv
Nitrogen deposition	Enhanced Dentener extrapolation (MsTMIP)	0.5-degree gridded product produced from the enhanced Deneter et al 2006 dataset (1860-2010) with constant 1860 levels extrapolated back to 850

Table 4. Environmental driver dataset. The environmental drivers for the Phase 1a intercomparison are annual CO2 concentration (global), land-use change and nitrogen deposition (extracted by site, dynamic through time), and soil type and biome (extracted by site, static in time).

Model Outputs

Model output should include annual output files, two decade-long snapshots of 3-hourly data, and a README file. Please see Table 5 below for the full list of required variables both for the entire time period and the decade-long snapshots. All variables should be output as **MONTHLY AVERAGES** unless otherwise indicated and packaged together into **100-year netCDF files** with dimensions of [time, lat, lon], which means that these variables will have dimensions of [1200, 30, 80]. Due to the size and number of output files, if a variable cannot be output by your model feel free to drop it from the file, but please note that in the README described below.

In addition to providing model output at a monthly time step, if possible we'd like to ask for **three decade-long 'snapshots' where we look at 3-hourly fluxes: 1600-1609, 1850-1859, and 1996-2005**. If your model does not have sub-daily resolution, please provide output at the highest native resolution. Please use the following naming: **[simulation acronym].[model acronym].[earliest year].nc** (e.g., PHA.ED2.1500.nc) where simulation acronym corresponds to the model descriptions in Table 1.
3-hourly (or the sub-daily timestep) output should be saved as monthly files using the naming convention:
[simulation acronym].[model acronym].[year].[month].nc (e.g., SPHA.ED2.1501.03.nc)

README file (REQUIRED!)

In addition to model outputs groups need to submit a README file that contains, at a minimum, the following information:

- **Model information**: Full model name and version; Model acronym; Preferred/standard citations; URL; Is the code used in this analysis publicly available? Where?
- **Modeling team:** Names of those who contributed significantly to producing the PALEON runs; Address; Email; Phone; Brief description of each person's contribution
- **Modifications:** Any model changes/tuning you made in order to produce your runs; Any additional input data sets your model required; List any of the standard input data sets or variables that you model did NOT require; Document any deviations from the PalEON protocol. **Please make particular note of any modifications to the meteorological data necessary to drive the model.** For example, ED2 internally uses linear interpolation of the 6-hourly drivers whereas other models use their own temporal downscaling approach and others still require aggregation up to daily or monthly time steps.
- **Traceability information**: any constant annual transfer rates within your model

for NPP allocation to plant carbon pools or between ecosystem carbon pools (Examples: fraction NPP allocated to wood, fraction slow SOM to passive SOM).

- **Model Settings:** Provide any settings files that detail the choices for model-specific conditions or settings as well as any available files that specify model parameters such as PFT characteristics or prescribed disturbance. (For example, the ED2IN files for ED2). If no settings file is available, please note the following settings in the ED2IN:
    - What modes of disturbance are used?
    - Is land use turned on or off?
    - Is there a nitrogen limitation scheme used in the runs?
    - Which PFTs are enabled?
- **Changes made from previously provided output:** This may include running the model with updated met driver, altered model settings, or changes in the post-processing.

We are asking for all of this up front so that we do not have to spend time tracking this all down later when it comes time to interpret model output and to write up results. This will also reduce 'surprises' related to authorship and allow us to know exactly who needs to be kept in the loop about analyses/presentations/manuscripts.

Table 5: List of required output variables. This variable list is kept updated at https://docs.google.com/spreadsheets/d/1f0LXVnHLglppztsFbYzDdaC5TMSZYUGg4OM__MnmaDM/edit?pli=1#gid=630592499. Since Phase 1a only includes site-level runs, all of the following variables should be output in the format: [time] or [time,classes].

Category	Variable name	Long name	Units	Sub-daily	Description
Diversity	PFT	PFT name	-	No	Name of each plant functional type or species included in the model. Dimensions: [PFT]
	Fcomp	Fractional Composition	kgC/kgC		**AGB** fractional composition of each PFT within each grid cell. Dimensions: [time, PFT]
	BA	Basal Area	m2/ha		Basal area by PFT Dimensions: [time, PFT]
	Dens	Stem density	1/ha		Stem Density by PFT Dimensions: [time, PFT]
	Estab	Establishment			New individuals Dimensions: [time, PFT]
	Mort	Mortality			Individuals lost through eath Dimensions: [time, PFT]

Category	Variable name	Long name	Units	Sub-daily	Description
Carbon Pools	AGB	Aboveground biomass	kgC/m^2	No	Total aboveground biomass
	TotLivBiom	Total living biomass			Total carbon content of living biomass (e.g. leaf +root+wood)
	TotSoilCarb	Total soil carbon			Total soil and litter cabon content over the entire soil profile
	CarbPools	Size of each carbon pool (e.g. leaf, wood, root, litter, CWD, soil)			However vegetation and soils are broken down in your model. Dimensions: [time,pool]
	poolname	Names of carbon pools			Names of each veg and soil carbon pool. Dim: [pool]
Carbon Fluxes	GPP	Gross primary productivity	kgC/m^2/s	Yes	
	AutoResp	Autotrophic Respiration			
	HeteroResp	Heterotrophic Respiration			
	NPP	Net Primary Productivity			NPP of each PFT within each grid cell. Dimensions: [time, PFT]
	NEE	Net Ecosystem Exchange			
	Fire	Fire Emissions			
	GWBI	Gross woody biomass increment	kgC/m2/month	No	Variable most analogous to tree-ring-derived change in stem biomass (before mortality/CWD flux) *moved from pools*
	CWDI	Coarse Woody Debris Increment			Variable most analogous to flux of woody material to the detrital pool resulting from mortality; corresponds to GWBI
	CPoolIn	Carbon flux into carbon pools	kgC/m^2/s	Yes	
	CPoolOut	Carbon flux out of carbon pools			
Energy fluxes	LW_albedo	Longwave Albedo	-	No	

Category	Variable name	Long name	Units	Sub-daily	Description
	SW_albedo	Shortwave Albedo	-		
	LWnet	Net Longwave Radiation	W/ m^2	Yes	
	SWnet	Net Shortwave Radiation			
	Qh	Sensible Heat			
	Qle	Latent Heat			
Other	LAI	Leaf Area Index	m^2/m^{-2}	Yes	
	Qs	Surface runoff	$kg/m^2/s$		
	Qsb	Subsurface runoff	$kg/m^2/s$		Drainage and subsurface lateral flow
	Evap	Total Evaporation			Sum of evaporative sources minus transpiration
	Tranp	Total Transpiration			Transpiration of each PFT within each grid cell. Dimensions: [time, PFT]
	SnowDepth	Total snow depth			
	SWE	Snow water equivalent	kg/m^2		Total water mass (ice plus liquid)
	SoilMoist	Soil moisture	kg/m^2		Soil water content in each model-defined soil layer. Dimensions: [time,lat,lon,nsoil]
	SoilTemp	Soil temperature	K		Soil temperature in each model-defined soil layer. Dimensions: [time,lat,lon,nsoil]
	SoilDepth	Soil layer depths	m	No	Depth to the bottom of each model-defined soil layer: Dimensions: [nsoil]
Met drivers	lwdown	Incoming long-wave radiation	W m-2	Yes	Incident longwave radiation averaged over the time step of the forcing data
	swdown	Incoming short-wave radiation	W m-2		Incident radiation in the shortwave part of the spectrum averaged over the time step of the forcing data
	precipf	Precipitation	kg m-2 s-1		The per unit area and time precipitation representing the sum of convective rainfall, stratiform rainfall,

Category	Variable name	Long name	Units	Sub-daily	Description
					and snowfall
	psurf	Surface pressure	Pa		Pressure at the surface
	qair	Specific humidity	kg kg-1		Specific humidity measured at the lowest level of the atmosphere
	tair	Air temperature	K		2 meter near surface air temperature
	wind	Wind speed	m s-1		Wind speed measured with a vertical coordinate in height of 10 m
	CO2	CO2 concentration	ppm		Carbon dioxide concentration in the air

Uploading Model Output

Model output should be uploaded to the "models" folder within "phase1a_model_output" on iPlant. If your model does not currently have a folder, please create one and within that location create a new folder titled [model name}.v1. If you have previously provided model output, please sequentially number your provided output according to the scheme listed above.

Please zip or tar each set of simulations into an archive with the following format: [simulation acronym]_[model name].zip/tar (i.e. PHA_ED2.tar, SPHA_ED2.tar, UPHA_ED2.tar, etc.). **Please send Christy a quick email notification (crollinson@gmail.com) after you've uploaded files**, so we can perform some quick checks on model output to make sure variables are provided (or documented as not possible from your model) and check for gross errors such as potential unit inconsistencies.

Name	**Email**	**Role**
Christy Rollinson	crollinson@gmail.com	PalEON modeling Postdoc (Boston U.)
Yao Liu	yaoliu@email.arizona.edu	PalEON modeling Postdoc (U. AZ)
Michael Dietze	dietze@bu.edu	PalEON modeling Co-PI (Boston U.)
Dave Moore	davidjpmoore@gmail.com	PalEON modeling Co-PI (U. AZ)
Jason McLachlan	jmclachl@nd.edu	PalEON lead PI (Notre Dame)

Table 5: PalEON Model-Data Inter-comparison Management Team.