Quality assessment (QA) tools of impact model output. An assessment of existing evaluation frameworks

Version 1.0 by Aris Koutroulis & Hannes Müller Schmied, 15.06.2021

Summary

Robust projections of future climate impacts rely on skillful modeling. To this aim, process based models become increasingly complex and so is the need for evaluation with objective comparisons against observations, for research and model improvement (Kumar et al. 2012; Hoffman et al. 2017).

The purpose of this document is to assess the existing evaluation frameworks that could facilitate the capacity building activities within PROCLIAS, fostering development, exchange and adoption of tools developed to benchmark the performance of process based models linked to ISIMIP and other individual modelling frameworks. In addition, this could help individual modelling teams to support their model development activities. Furthermore, adopting existing QA tools to an automatic framework for the models used in ISIMIP might lead to a model verification/model benchmark protocol for impact models through the umbrella of PROCLIAS.

1 Quality assessment (QA) tools – an overview

Several tools for model evaluation and benchmarking of land surface models are available.

- The International Land Model Benchmarking (ILAMB) Project has developed initial prototype benchmarking system (Collier et al. 2018) to assess and improve the performance of land models through international cooperation and to inform the design of new measurement campaigns and field studies to reduce uncertainties associated with key earth system processes and feedbacks.
- The ESMValTool, a community diagnostics and performance metrics tool designed to improve comprehensive and routine evaluation of Earth system models participating in the Coupled Model Intercomparison Project (Eyring et al. 2020).
- The Protocol for the Analysis for Land Surface models (PALS), an online web application for the automated evaluation and benchmarking of land surface model (LSM) simulations (Best et al. 2015).
- The PCMDI Metrics Package (PMP) provides objective comparisons between Earth System Models (ESMs) and available observations (Gleckler et al. 2016).
- The NASA Land Surface Verification Toolkit (LVT) is a system for land surface model evaluation and benchmarking by comparing them with data from observational networks, remote-sensing platforms and similar estimates from other modeling frameworks (Kumar et al. 2012).

Table 1: Overview with key characteristics of the tools evaluated

Tool	MIP	Code development possible?	Open source?	Topic	Reference data extendable?
ILAMB	C-LAMP	yes	yes	multiple	yes
EMSEvalTool	CMIP	yes	yes	multiple	yes
PALS		No		atmosphere	
PMP	CMIP	No	yes	atmosphere, climatology	
LVT	LIS	No (?)	yes	hydrology	

In this report we focus on the ILAMB and ESMValTool systems but also describe briefly the other evaluation frameworks.

1.1 The International Land Model Benchmarking (ILAMB) System

ILAMP is an open source model benchmarking software package that generates graphical diagnostics and scores model performance in support of the International Land Model Benchmarking (ILAMB) project (Collier et al. 2018), leveraging prior work on the Carbon- Land Model Intercomparison Project (C-LAMP) (Hoffman et al. 2017). It assesses model performance for variables in categories of biogeochemistry, hydrology, radiation and energy and climate forcing (Table 1). Additional datasets may be added to the sample benchmark comparison. For each of these variables, the packages generate graphical diagnostics (Figure 1) and score model performance for the period mean over whole years and its bias, RMSE, spatial distribution, interannual coefficient of variation, and seasonal cycle and long-term trend.

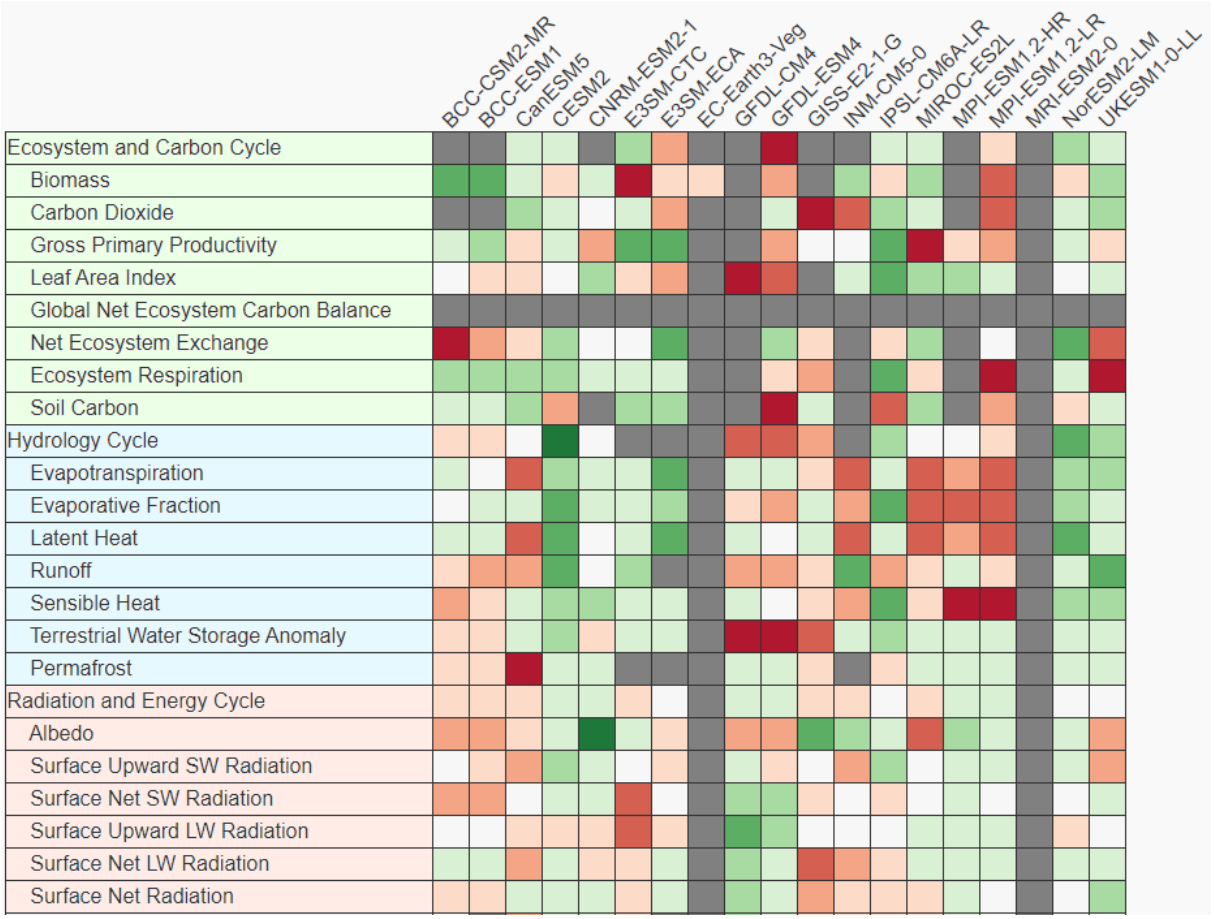

***Figure 1**: Coupled Model Intercomparison Project Phase 6 (CMIP6) Benchmarking preliminary results (https://www.ilamb.org/results/)*

The ILAMB benchmarking software is written in Python and depends on a few packages which extend the language's usefulness in scientific applications. ILAMP (current version 2.6, May 2021) is continuously developing by e.g., enriching with new and updating existing observational datasets to their most current versions, adding case specific metrics and new ways of illustrating scoring outputs.

***Table 2**: This table shows the benchmarks and data sources in three topical areas of ILAMB. Grouped information from tables 2, 3 and 4, adopted by Collier et al. (2018). Information on certainty, scale and process are detailed in Table 2. Reference sources are included in Collier et al. (2018).*

	Ecosystem and Carbon Cycle								Hydrology Cycle						Radiation and Energy					
	Biomass	Burned area	Gross primary productivity	Leaf area index	Global net ecosystem carbon balance	Net ecosystem exchange	Ecosystem respiration	Soil carbon	Evapotranspiration	Evaporative fraction	Latent heat	Runoff	Sensible heat	Terrestrial water storage anomaly	Albedo	Surface upward SW radiation	Surface net SW radiation	Surface upward LW radiation	Surface net LW radiation	Surface net radiation
PROCESS:	5	4	5	3	5	5	4	5	5	5	5	5	2	5	1	1	1	1	1	2
DATASET	CERTAINTY								CERTAINTY						CERTAINTY					
Tropical (Saatchi et al., 2011)	4																			
NBCD2000 (Kellndorfer et al., 2013)	4																			
USForest (Blackard et al., 2008)	4																			
GFED4S (Giglio et al., 2010)		4																		
Fluxnet (Lasslop et al., 2010)			3			3	2				3		3							4
GBAF (Jung et al., 2010)			3			2	2			3	3		3							
AVHRR (Myneni et al., 1997)				3																
MODIS (De Kauwe et al., 2011)				3					3						4					
GCP (Le Quéré et al., 2016)					4															
Hoffman (Hoffman et al., 2014)					4															
HWSD (Todd-Brown et al., 2013)								3												
NCSCDV22 (Hugelius et al., 2013)								3												
GLEAM (Miralles et al., 2011)									3											
Dai (Dai & Trenberth, 2002)												3								
GRACE (Swenson & Wahr, 2006)														5						
CERES (Kato et al., 2013)															4	4	4	4	4	4
GEWEX.SRB (Stackhouse et al., 2011)															4	4	4	4	4	4
WRMC.BSRN (König-Langlo et al., 2013)																4	4	4	4	4

***Table 2**: The ILAMB Rubric Used to Assign Relative Weights of a Data Set, adopted by Collier et al. (2018)*

Score	CERTAINTY	SCALE	PROCESS
1	No given uncertainty, significant methodological issues affecting quality	Site level observations with limited space/time coverage	Observations that have limited influence on the targeted Earth system dynamics
2	No given uncertainty, some methodological issues affecting quality	Partial regional coverage, up to 1 year	Observations have direct influence on the targeted Earth system dynamics

3	No given uncertainty, methodology has some peer review	Regional coverage, at least 1 year	Observations useful to constrain processes that contribute to the targeted Earth system dynamics
4	Qualitative uncertainty, methodology accepted	Important regional coverage, at least 1 year	Observations well suited to constrain important processes
5	Well-defined and relatively low uncertainty	Global scale spanning multiple years	Observations well suited to constrain important processes
Note. A score for each data set is assigned in each of three areas. These scores are then combined multiplicatively and used to determine relative importance for a data set with respect to a given variable. ILAMB = International Land Model Benchmarking.			

1.2 The Earth System Model Evaluation Tool (ESMValTool) v2.0

The ESMValTool is a community tool created for routine evaluation of ESMs participating in the Coupled Model Intercomparison Project (CMIP). It contains the core functionalities (ESMValCore, e.g. to prepare the CMIP data) as a Python package and a diagnostic part with metrics, scientific applications and diagnostics, called recipes and translated with YAML (allowing the support of different programming languages for the diagnostics) and the workflow manager to the ESMValCore. The ultimate goal is to embed ESMValTool alongside the Earth System Grid Federation as part of a routine evaluation of CMIP model simulation. Some of the main features are listed at the website https://www.esmvaltool.org/about.html highlighting (among other) the flexibility of this tool. In general, this tool allow diagnostics of model simulations against observations, against other models or to compare different versions of the same model. Descriptions of ESMValTool are available at (Righi et al. 2020; Eyring et al. 2020; Lauer et al. 2020; Weigel et al. 2021). The general flow chart is visible in Figure 2

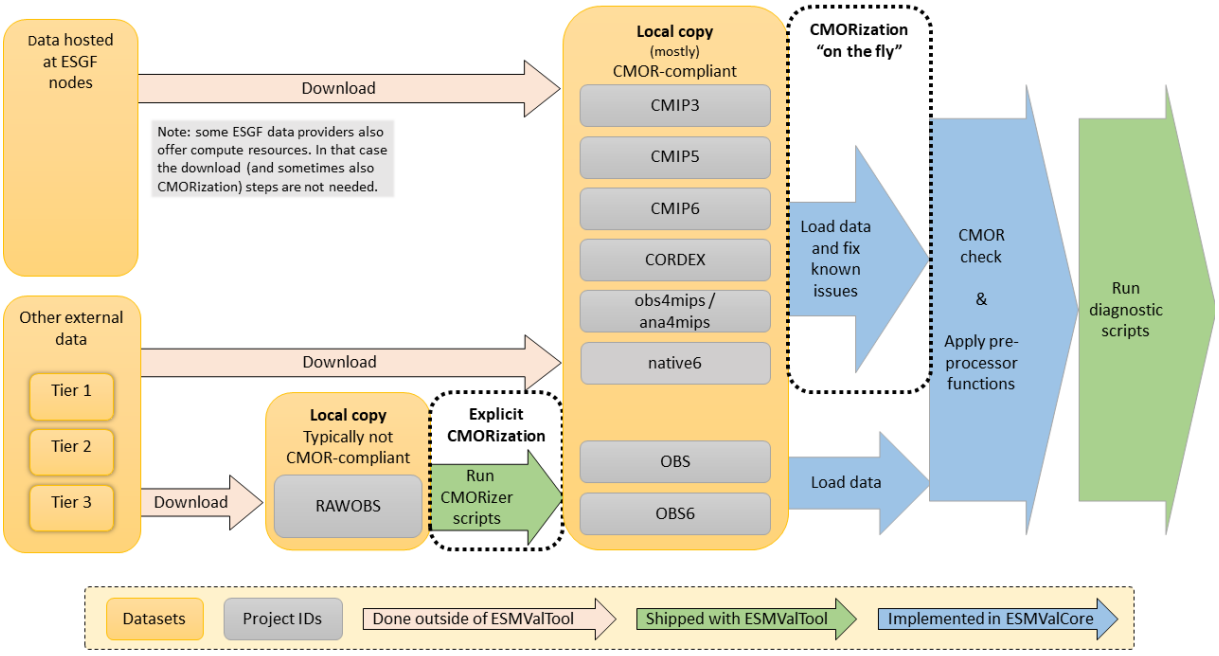

***Figure 2**: Workflow and tools used in EMSValCore/EMSValTool (*https://esmvalgroup.github.io/ESMValTool_Tutorial/*)*

ESMVal is continuously developed since 2016 and provides, next to the general tools and frameworks also the possibility to inform the public with a tool from Freie University Berlin (https://cmip-esmvaltool.dkrz.de/).

1.3 The Protocol for the Analysis of Land Surface models (PALS) and its successor modelevaluation.org

PALS was an online platform that enables comparison of land surface model outputs to site-based flux tower data (Best et al. 2015). Ist successor https://modelevaluation.org/ is a web application for evaluating and benchmarking models (Figure 3). Currently, station-, catchment-, regional- and globalscale data sets are included such as MODIS evaporation, biomass or albedo. Maintained by the University New South Wales.

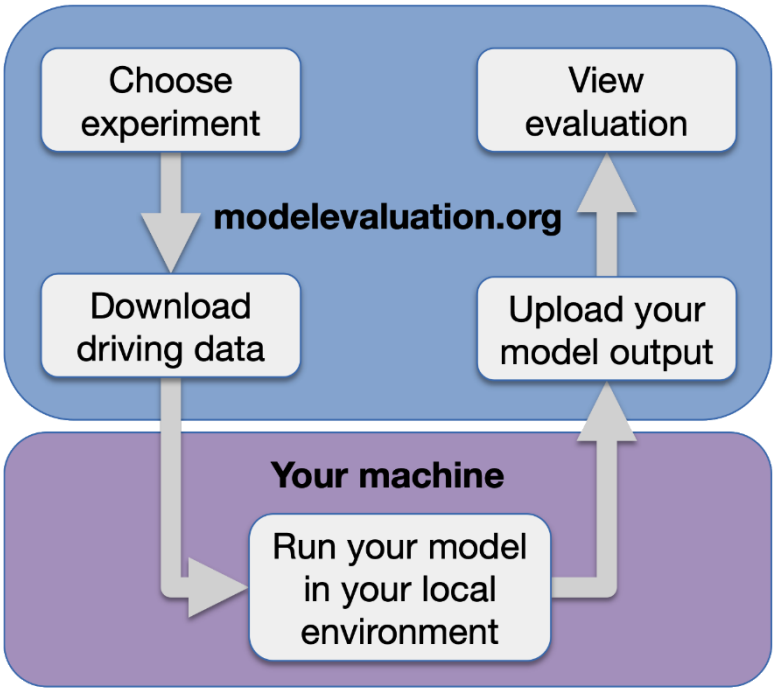

***Figure 3**: Schematic of modelevaluation.org (from their website)*

1.4 PCMDI Metrics Package

The PCMDI metrics package (PMP) is designed as "quick look" comparisons of ESMs with each other and observations. It is produced for CMIP6 (and earlier) model phases (Gleckler et al. 2016). The PMP creates of four parts: 1) the analysis software, 2) an observationally-based database of global time series and climatologies (mainly reanalyses), 3) a database of performance metrics and 4) documentation and demos (https://github.com/PCMDI/pcmdi_metrics). Focus is on precipitation and climatologic features such as ENSO, MJO and monsoon characteristics.

1.5 The NASA Land Surface Verification Toolkit (LVT)

The LVT is designed to evaluate, analyze, compare and benchmark the outputs of the Land Information System LIS (https://lis.gsfc.nasa.gov). It is designed predominantly for terrestrial hydrology datasets and can handle multiple reference data sets and types. Figure 4 shows the schematic of LVT in its framework.

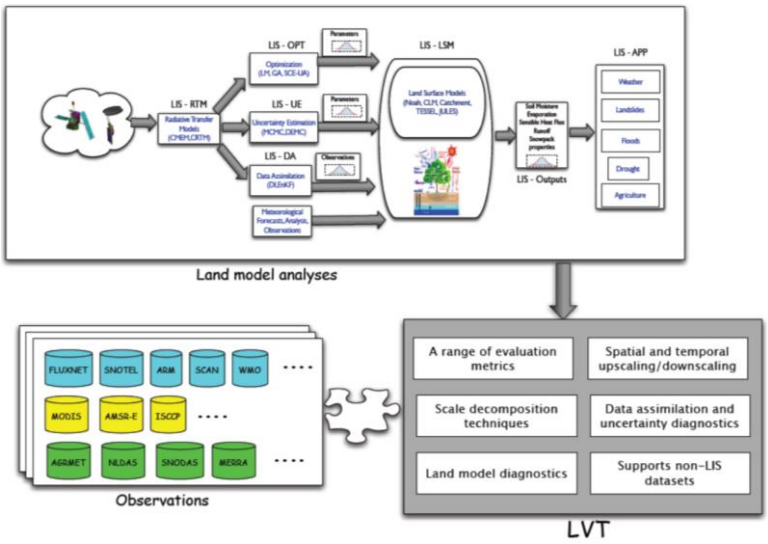

***Figure 4**: Schematic of LVT and its connection to LIS* (Kumar et al. 2012)

2 Outlook

The main outlook questions can be defined as:

1) Can such kind of tools be considered as universal tools for all ISIMIP sectors?
2) Which tools are adaptable for the ISIMIP needs?
3) How can we allocate resources to test those tools with real ISIMIP data?
4) Is it feasible to start evaluating those tools with one or two sectors to transmit the experiences and information to all sectors at a later stage?

References

Best MJ, Abramowitz G, Johnson HR, et al (2015) The plumbing of land surface models: Benchmarking model performance. J Hydrometeorol 16:1425–1442. https://doi.org/10.1175/JHM-D-14-0158.1

Collier N, Hoffman FM, Lawrence DM, et al (2018) The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation. J Adv Model Earth Syst 10:2731–2754. https://doi.org/10.1029/2018MS001354

Eyring V, Bock L, Lauer A, et al (2020) Earth System Model Evaluation Tool (ESMValTool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP. Geosci Model Dev 13:3383–3438. https://doi.org/10.5194/gmd-13-3383-2020

Gleckler PJ, Doutriaux C, Durack PJ, et al (2016) A more powerful reality test for climate models. Eos (United States) 97:20–24. https://doi.org/10.1029/2016eo051663

Hoffman FM, Koven CD, Keppel-Aleks G, et al (2017) International Land Model Benchmarking (ILAMB) 2016 Workshop Report, DOE/SC-0186. Germantown, Maryland, USA

Kumar S V., Peters-Lidard CD, Santanello J, et al (2012) Land surface Verification Toolkit (LVT) - A generalized framework for land surface model evaluation. Geosci Model Dev 5:869–886. https://doi.org/10.5194/gmd-5-869-2012

Lauer A, Eyring V, Bellprat O, et al (2020) Earth System Model Evaluation Tool (ESMValTool) v2.0 – diagnostics for emergent constraints and future projections from Earth system models in CMIP. Geosci Model Dev 13:4205–4228. https://doi.org/10.5194/gmd-13-4205-2020

Righi M, Andela B, Eyring V, et al (2020) Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview. Geosci Model Dev 13:1179–1199. https://doi.org/10.5194/gmd-13-1179-2020

Weigel K, Bock L, Gier BK, et al (2021) Earth System Model Evaluation Tool (ESMValTool) v2.0 – diagnostics for extreme events, regional and impact evaluation, and analysis of Earth system models in CMIP. Geosci Model Dev 14:3159–3184. https://doi.org/10.5194/gmd-14-3159-2021