https://doi.org/10.5194/gmd-2023-188
Preprint. Discussion started: 23 October 2023

The computational and energy cost of simulation and storage for climate science: lessons from CMIP6

Mario C. Acosta[1], Sergi Palomas[1], Stella Paronuzzi[1], Jean-Claude Andre[2], Joachim Biercamp[3], Pierre-Antoine Bretonniere[1], Reinhard Budich[4], Miguel Castrillo[1], Arnaud Caubel[5], Francisco Doblas-Reyes[1], Italo Epicoco[6], Uwe Fladrich[7], Alok Kumar Gupta[8], Bryan Lawrence[9], Philippe Le Sager[10], Grenville Lister[9], Marie-Pierre Moine[11], Jean-Christophe Rioual[12], Joussame Sylvie[5], Sophie Valcke[11], Niki Zadeh[13], and Venkatramani Balaji[14]

[1]Barcelona Supercomputing Center, Plaça d'Eusebi Güell, 1-3, 08034 Barcelona, Spain
[2]French National Centre for Scientific Research, 3 Rue Michel Ange, 75016 Paris, France
[3]German Climate Computing Centre, Bundesstraße 45a, 20146 Hamburg, Germany
[4]Max Planck Institute, Hofgartenstr. 8, 80539 Munich, Germany
[5]Institut Pierre-Simon Laplace, 11 Bd d'Alembert, 78280 Guyancourt, France
[6]Euro-Mediterranean Center on Climate Change, Via della Libertà, 12, 30121 Venezia VE, Italy
[7]Swedish Meteorological and Hydrological Institute, SE-601 76 Norrköping, Sweden
[8]Norwegian Research Centre, Nygårdsgaten 112, 5008 Bergen, Norway
[9]National Centre for Atmospheric Science, Fairbairn House, 71-75 Clarendon Rd, Woodhouse, Leeds LS2 9PH, United Kingdom
[10]Royal Netherlands Meteorological Institute, Utrechtseweg 297. NL-3731 GA De Bilt, Netherlands
[11]European Center for Advanced Research and Training in Scientific Computing, 42 Av. Gaspard Coriolis, 31100 Toulouse, France
[12]Meteorological Office, Fitzroy road, Exeter, Devon, EX1 3PB, United Kingdom
[13]National Oceanic and Atmospheric Administration, 1401 Constitution Avenue NW, Room 5128, Washington, DC 20230, USA
[14]High Meadows Environmental Institute, Princeton University, Guyot Hall, Room 129, Princeton, NJ 08544-1003, USA

Correspondence: Mario C. Acosta (mario.acosta@bsc.es), Sergi Palomas (sergi.palomas@bsc.es)

Abstract. The Coupled Model Intercomparison Project (CMIP) is one of the biggest international efforts to better understand past, present and future climate changes in a multi-model context. A total of 21 Model Intercomparison Projects (MIPs) were endorsed in its 6th phase (CMIP6), which included 190 different experiments that were used to simulate 40000 years and produced around 40 PB of data in total. This paper shows the main results obtained from the collection of performance metrics done for CMIP6 (CPMIP). The document provides the list of partners involved, the CPMIP metrics per institution/model, and the approach used for the collection and the coordination behind this process. Furthermore, a section has been included to analyze the results and prove the usefulness of the metrics for the community. Moreover, we describe the main difficulties faced during the collection and propose recommendations for future exercises.

1 Introduction

Earth System Models (ESMs) are an essential tool for understanding the Earth's climate and the consequences of climate change, which are crucial to the design of response policies to address the current climate emergency resulting from anthropogenic emissions. Modelling the Earth is inherently complex. ESMs are among the most challenging applications that the HPC industry has had to face, requiring the most powerful computers available, consuming vast amounts of energy in computer power, and producing massive amounts of data in the process (Wang and Yuan 2020; Wang et al. 2010; Fuhrer et al. 2014; McGuffie and Henderson-Sellers 2001; Dennis et al. 2012). In such a heterogeneous context -where multiple large modelling frameworks, HPC machines and institutions coexist- running coordinated experiments is key to achieving the most accurate climate science projections. The Coupled Model Intercomparison Project Phase 6 (CMIP6) is noticeably one of the best examples of a multi-model project led by the World Climate Research Programme (WCRP), in which 124 different model configurations from 45 centres[1] have participated. As seen by Eyring et al. (2016), having multiple institutions collaborating on such a wide and diverse project has made it possible to better understand fundamental elements of the Earth system, such as how it responds to forcing, the consequences of systematic model biases and assessing future climate changes given the variability, predictability and uncertainties that are still present in current ESMs.

Virtually all models are designed to exploit the parallelism of HPC machines so that we can get the results in a reasonable amount of time while trying to make the best use of the HPC platform. While the technology underneath keeps improving every year (in Petaflops/s, memory bandwidth, I/O speed, etc.) the software evolves much more slowly, and one of the endless and most important challenges that Earth system modelling faces is how to make use of the HPC platforms effectively (Bauer et al. 2021; Hill et al. 2004; Giles and Reguly 2014; Takizawa et al. 2013). Enhancing the performance of these models is crucial to boost the rate at which they can grow (in the resolution, complexity and features simulated) and to allow running bigger ensemble sizes to minimize the model's inherent uncertainty (Evensen 2003; Evans et al. 2013; Du et al. 2020).

The performance of ESMs is hardly limited only by one but by multiple bottlenecks that depend on the model itself and on the properties of the HPC platform on which they run, for instance: models using higher resolutions may benefit from (or be limited by) the speed of the network as the data is split into many nodes, memory-bound models will benefit from having more memory available per core and with faster transmission speed while compute-bound models will perform better in faster CPUs, models that produce more output will run faster on infrastructures with higher capacities for I/O operations, models that include more individual components will be limited by the load-balance achieved between them and by the coupler performance, etc. Balaji et al. (2017) proposed a set of 12 performance metrics, the Computation Performance for Model Intercomparison Project (CPMIP), at the inception of CMIP6 which take into account the structure of ESMs and how they are executed in real experiments. Thus, they include the model and platform properties, the computational cost (speed, cost, and energy), measures for the coupling and IO overhead, and for the memory consumption. Each one is described in detail in Tab. 1.

In this paper, we present in Sec. 2 the collection of CPMIP metrics done for the IS-ENES3 project (Joussaume, 2010), which accounted for 30 different models/configurations that were used to simulate almost 500 000 years during the CMIP6 exercise

[1] http://esgf-ui.cmcc.it/esgf-dashboard-ui/data-archiveCMIP6.html, retrieved October 23, 2023.

https://doi.org/10.5194/gmd-2023-188
Preprint. Discussion started: 23 October 2023

Table 1. List of CPMIP metrics collected

Metric	Used to evaluate
Resolution (Resol)	number of grid points NXxNYxNZ per component
Complexity (Cmplx)	number of prognostic variables per component
Platform	machine measurements: core count, clock frequency, and double-precision op. per clock cycle
Simulation Years Per Day (SYPD)	number of simulated years per day (24h) of execution time
Core-hours per Simulated Year (CHSY)	cost in core-hours per simulated year
Actual SYPD (ASYPD)	how queue time and interruptions affect the complete experiment duration
Parallelization (Paral)	total number of cores allocated for the run
Joules Per Simulated Year (JPSY)	energy needed per year of simulation
Memory Bloat (Mem B)	ratio between actual and ideal memory size
Data Output cost (DO)	computing cost for performing I/O
Data Intensity (DI)	amount of data produced after 1 year of simulation divided by the CHSY
Coupling Cost (Cpl C)	computing cost of the coupling algorithm and load imbalance

on 14 different HPC machines and by 9 independent institutions. All experiments are listed in Tab. 2, with the institution in charge, the model name, HPC platform, ocean (OCN) and atmosphere (ATM) resolutions and the main reference to the experiment configuration. In addition, Tab. 3 shows the complete collection of CPMIP metrics for each one of the models, and Tab. 4 shows the list of HPC machines that have been used to run these models. Furthermore, in Sec. 3 we include the analysis of the results to highlight the most important insights we can get from this data and to discuss what are the strengths and difficulties that we have encountered and what ought to be improved for future inter-model performance comparisons.

2 Data collection

The collection process was coordinated and supervised to get the metric results, including meetings, reporting, and surveys at different moments of the CMIP6 simulations (before, during, and after the simulation runs). All the partners included in Tab. 2 were invited to participate in the tracking process. The coordination, meetings, and reporting were useful to evaluate the state of the collection from the partners, and support was provided to those institutions that required it during the collection process.

The collection was divided into two steps: the first comprehends the collection up to March 2020, coinciding with the first IS-ENES3 general assembly where the first results were presented; the second includes the data collected up to the end of 2020 when all the institutions had finished the CMIP6 runs. Finally, IS-ENES3 completed the last update to the Earth System Documentation [2] (ES-DOC) in the middle of 2021, publishing CPMIP along with the other CMIP6 results.

As the reader can see, not all institutions provided the full collection of CPMIPs. The metrics frequently missing are the *Cpl C*, *Mem B*, and *DO*, primarily due to the challenges involved in their collection compared to metrics like *SYPD* or Par-

[2] https://es-doc.org/

https://doi.org/10.5194/gmd-2023-188
Preprint. Discussion started: 23 October 2023

Table 2. List of Institutions with the model, HPC platform, and resolution used for the ATM and OCN components

Institution	Model	HPC machine	ATM resol	OCN resol	Ref
BSC	EC-Earth3	MareNostrum4	0.7	1.0	Döscher et al. (2022)
	EC-EarthVeg		0.7	1.0	
CMCC	CM2-SR5	Zeus	1.0	1.0	Lovato et al. (2022)
CNRM-CERFACS	CNRM-CM6-1-atm	Beaufix2	1.4		Voldoire et al. (2019)
	CNRM-CM6-1		1.4	1.0	
	CNRM-CM6-1-HR-atm		0.5		
	CNRM-CM6-1-HR		0.5	1/4	
	CNRM-ESM2-1-atm		1.4		Séférian et al. (2019)
	CNRM-ESM2-1		1.4	1.0	
DKRZ	MPI-ESM1-HR	Mistral	1.0	0.4	Müller et al. (2018)
GFDL	OM4-p5	Gaea		0.5	Dunne et al. (2020)
	ESM4-piC		1.0	0.5	
	CM4-piC		1.0	1/4	
	OM4-p25			1/4	
IITM	IITM-ESM	Intel AADITYA	1.875	1.0	Krishnan et al. (2021)
IMPE	BESM	xc50	1.875	1.0	Veiga et al. (2019)
IPSL	IPSL-CM6A	Irene-SKL/Curie	2.5	1.0	Boucher et al. (2020)
KNMI	EC-Earth3	Rhino	0.7	1.0	Döscher et al. (2022)
	EC-Earth3-AerChem		0.7	1.0	
MPI	MPI-ESM1-LR-ATM	Mistral	4.0		Müller et al. (2018)
	MPI-ESM1-LR-LAND				
	MPI-ESM1-LR		1.875	1.5	
NERC	UKESM1-AMIP	Archer xc30	4.0		Sellar et al. (2020)
	UKESM1-0-LL		1.875	1.0	
	HadGEM3-GC3.1-LL		1.875	1.0	Williams et al. (2018),
	HadGEM3-GC3.1-HM		0.8	1/12	
NorESM2	NorESM2-LM	Fram	2.5	1.0	Seland et al. (2020)
	NorESM2-MM		1.0	1.0	
SMHI	EC-EarthVeg	Tetralith/Beskow	0.7	1.0	Döscher et al. (2022)
UKMO	UKESM1-0-LL	xce xc40	1.875	1.0	Sellar et al. (2020)
	HadGEM3-GC3.1-LL		1.875	1.0	Williams et al. (2018)
	HadGEM3-GC3.1-MM		0.8	1/4	

allelization, which are relatively easier to obtain. While certain institutions required support to generate the necessary data, three institutions were unable to produce these measures entirely due to constraints related to time, resources during or after the CMIP6 runs, and in some cases, changes in the underlying computational infrastructure.

https://doi.org/10.5194/gmd-2023-188
Preprint. Discussion started: 23 October 2023

Geoscientific Model Development
Discussions
Open Access

Table 3. List of Institutions with the model and CPMIP metrics

Institution	Model	Resol	Cmplx	SYPD	ASYPD	CHSY	Paral	JPSY	Cpl C	Mem B	DO	DI	Useful SY
BSC	EC-Earth3	1.99E+07	34	15.2	9.87	1213	768	4.41E+07	0.08	59.5	1.12	0.041	14020
	EC-EarthVeg	1.99E+07		12.36	7.42	1491	768	4.87E+07	0.1	68.48	1.13	0.059	252
CMCC	CM2-SR5	6.94E+06	397	6.68	6.5	2069	576	1.67E+09	0.074	17.8	1.04	0.05	965
CNRM-CERFACS	CNRM-CM6-1-atm	2.98E+06	128	7.3	6.1	1292	393	3.50E+07					5723
	CNRM-CM6-1	1.02E+07	181	8.1	7.3	1352	400	3.38E+07					22241
	CNRM-CM6-1-HR-atm	2.36E+07	128	2.2	1.8	1541	520	4.80E+07					1190
	CNRM-CM6-1-HR	1.37E+08	181	1.5	1.48	4289	840	1.07E+08					1642
	ESM2-1-atm	2.98E+06	335	7.1	6.4	8520	781	2.28E+08					1759
	ESM2-1	1.10E+07	393	4.7	4.4	21552	1347	5.28E+08					11761
DKRZ	MPI-ESM1-HR	2.00E+07		13.33	11	4710	2616	3.21E+08					1864
GFDL	OM4-p5	3.32E+07	13	15.9	12.22	1962	1300	7.50E+07	0.14	33.61		0.039	300
	ESM4-piC	3.76E+07	140	8.65	7.46	13576	4893	5.19E+08	0.27	40.57		0.003	1124
	CM4-piC	1.28E+08	31	9.98	8.16	15388	6399	3.72E+08	0.13	47.64		0.018	657
	OM4-p25	1.26E+08	11	11.5	7.05	9748	4671	5.88E+08	0.26	16.09		0.006	300
IITM	IESM	1.83E+06	168	8	7	996	332	3.81E+07		36.7			845
IMPE	BESM	6.88E+06	132	3.6	3.4	1853	278					0.02	360
IPSL	IPSL-CM6A	1.06E+07	750	12	11.5	1900	950	1.16E+08	0.05	10	1.2	0.070	53000
KNMI	EC-Earth3	1.99E+07	34	16.2	16.2	1286	868						1009
	EC-Earth3-AerChem	2.06E+07		3.03	3.03	3549	448						730
MPI	MPI-ESM1-LR-ATM	8.66E+05		45.9	25.2	163	312	1.11E+07					991
	MPI-ESM1-LR-LAND	8.33E+05		282.8	265.4	3	36	1.39E+05					2460
	MPI-ESM1-LR	3.12E+06		55.6	22.7	379	878	2.56E+07					18860
NERC	UKESM1-AMIP	2.35E+06	202	1.64	1.41	7376	504	1.04E+08		52.5	1.31	0.003	45
	UKESM1-0-LL	1.14E+07	252	2.02	1.1	8554	720	3.18E+08	0.078	28	1.19	0.005	195
	HadGEM3-GC3.1-LL	1.14E+07	150	4.25	1.06	12198	2160	4.33E+08	0.047	56.8	1.41	0.016	70
	HadGEM3-GC3.1-HM	1.99E+08	54	0.58	0.46	192662	4656	7.70E+09	0.21	154		0.001	65
	HadGEM3-GC3.1-HH	1.26E+09	54	0.49	0.34	588931	12024	2.30E+10		183	1.41	0.0004	65
NorESM2	NorESM2-LM	1.01E+07		13.84	3.03	1665	960	5.60E+07	0.035			0.065	5463
	NorESM2-MM	1.14E+07		8.96	6.14	4886	1824	1.65E+08	0.32			0.060	1021
SMHI	EC-EarthVeg	1.99E+07		12.44	6.65	1667	864					0.028	6337
UKMO	HadGEM3-GC3.1-LL	1.14E+07	228	4	3.55	13392	2232	4.97E+08	0.061	46	1.03	0.074	5610
	UKESM1-0-LL	1.14E+07	372	4.3	3.6	16074	2880	5.97E+08	0.098	4.6	1.03	0.019	15435
	HadGEM3-GC3.1-MM	1.44E+08	236	1.65	1.32	62836	4320	2.33E+09	0.105	120	1.02	0.050	2386

Table 4. List of HPC machines

Machine	total cores	cores per node	Mem node (GB)	Mem per core (GB)	network	CPU family	CPU freq (GHz)	Rpeak (PFlop/s)	Linpack (PFlop/s)	Power (kW)	HPCG (TFlop/s)	PUE
UKMO xc40	241920	36	192	5.33	Aries Interconnect	E5 Broadwell	2.1	8.13	7.04			1.35
BSC-MN4	155520	48	96	2.00	Intel Omni-Path	Platinum Skylake	2.1	10.3	6.22	1632	122.24	1.35
NERC-Archer xc30	118080	24	64	2.67	Aries Interconnect	E5 Ivy	2.7	2.55	1.64		80.79	1.1
DKRZ-MPI-Mistral	100200	30	68	2.25	InfiniBand	E5 Haswell	2.29	3.96	3.01	1116	44.11	1.19
IPSL-Curie	80640	16	64	4.00	InfiniBand	E5 Sandy Bridge	2.7	1.67	1.36	2132	50.99	1.43
IPSL-Irene	79488	48	192	4.00	InfiniBand	Platinum Skylake	2.7	6.64	4.07	917	52.68	
CNRM-CERFACS-beaufix2	73440	40	64	1.60	InfiniBand	E5 Broadwell	2.2	2.59	2.16	830	35.34	
PDC-Beskow	65920	32	64	2.00	Aries Interconnect	E5 Haswell	2.3	2.44	1.8	842		
NSC-Tetralith	61056	32	96	3.00	Intel Omni-Path	Gold Skylake	2.1	4.34	2.97		65.24	
IITM-AADITYA	38144	16	64	4.00	InfiniBand	E5 Haswell	2.6	0.79	0.72	790		
NorESM-Fram	32256	32	64	2.00	InfiniBand	E5 Broadwell	2.1	1.1	0.95			
KNMI-Rhino	4752	12	24	2.00	InfiniBand	Nehalem	3.06	0.058				
INPE-xc50	4080	40	192	4.80	Aries Interconnect	Gold Skylake	2.4	0.313				
CMCC-Zeus	12528	36	96	2.67	InfiniBand	Gold Skylake	3	1.202				1.84

2.1 Additional data collected

The CPMIP metrics serve not only as a means of computational evaluation but also provide valuable insights for broader analysis. In light of this, we collaborated with the Carbon Footprint Group created within the IS-ENES3 project, which was responsible for evaluating the Total Energy Cost associated with the CMIP6 experiments, enabling us to provide the first estimation of the carbon footprint related to those experiments. The Total Energy Cost of an experiment is the product of the useful simulated years, defined as years of simulation that produced data with scientific value that was either shared between the groups or kept within the producer group for scientific analysis, and the *JPSY*. The Carbon Footprint was calculated following Eq. 1.

$$\textit{Carbon Footprint} = \textit{Total Energy Cost} \times \textit{CF} \times \textit{PUE} \tag{1}$$

where the *Total Energy Cost* is the cost of running the whole experiment in MWh, *CF* is the greenhouse gas conversion factor from MWh to CO2 kilogram according to the supplier bill or the country energy mix, and *PUE* (*Power Usage Effectiveness*) accounts for other costs sustained from the data-center, such as cooling. The results for all the institutions that participated in the study during the CPMIP collection are shown during the analysis section in Tab. 10.

3 Analysis

Analyzing metrics derived from diverse models, executed on multiple platforms, and managed by independent institutions presents a non-trivial challenge. Moreover, the presence of missing values further complicates the analysis, making it difficult to substitute them with estimations or interpolations, particularly given the relatively limited size of the dataset.

Our approach consisted of first: validating the metrics provided by the institutions. We have sometimes found that the metrics reported for some models were orders of magnitude apart from the rest. In this case, we started actively communicating with the institutions asking them to double-check the values and assisting them in the re-computation process. After going through this process for each one of the metrics and models we came up with the values reported in Sec. 2: in Tab. 2 and Tab. 3 the reader can find the complete list of models for which the CPMIP metrics were collected, with the name of the institution that was in charge for the run, the resolution used for the OCN and ATM, the reference for the experiment configuration, and the CPMIP metrics. Additionally, we include in Tab. 4 the most relevant information on the HPC platforms used by the institutions and some supplementary metrics in Tab. 10 related to the execution costs in CO2 emissions.

Later, and for each of the metrics analyzed in detail in the following sections, we filtered by model selecting those where the metric was provided, sorting and/or grouping them by the reported value. Finally, to uncover possible relations among the metrics, we have used both statistical approaches (e.g. Pearson's correlation, Freedman et al., 2007) and qualitative analysis.

https://doi.org/10.5194/gmd-2023-188
Preprint. Discussion started: 23 October 2023

3.1 Resolution

The first attempt to extract valuable information from Tab. 3 was done by grouping the experiments by resolution, since for the moment we want to compare the performance achieved by ESMs whose target is similar. We are ignoring here the fact that for some simulations the set-up has fewer grid points (e.g. reduced Gaussian in the atmosphere or removal of land points in the ocean) and we are using the total size of the corresponding regular grid. The resolution of a component is measured as the number of grid points it has (NX x NY x NZ), and the total resolution is given by the sum of the resolutions of their constituents. There is not a strict consensus on the connection between the number of grid points and the categorization of low, medium, and high resolutions. Thus, for the grouping we have used both the naming provided by the institution in charge of the experiment and the total number of grid points used for each model configuration. Most configurations have been categorised as low resolution and use up to 2.10E+07 grid-points in total (see Fig. 1), and only those with an OCN resolution under 0.5 degree are treated as medium-high resolution configurations (see Fig. 2).

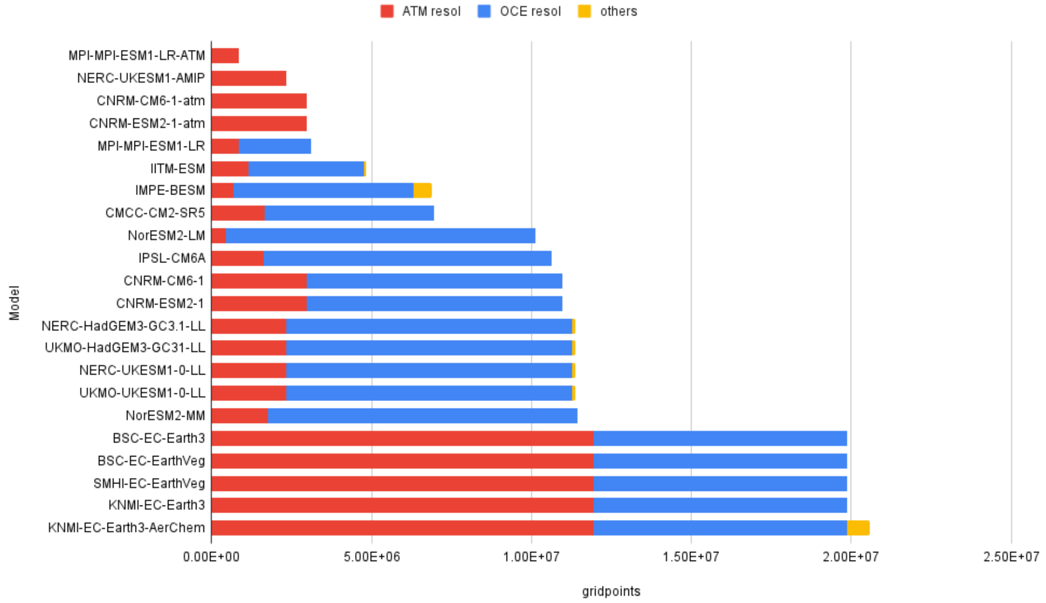

Figure 1. Atmosphere and ocean grid-points for low-resolution experiments. The yellow color refers to components that are contributing to the atmosphere or the ocean but can not be counted as a general circulation model per se (e.g. land-surface, sea ice, vegetation, etc.).

We see in Fig. 1 the low-resolution experiments. The number of grid points for the ATM (red) and OCN (blue) components for each model/institution has been listed in ascending order. We observe that except for EC-Earth, all other models run the OCN at a higher resolution than the ATM. More precisely, the OCN resolution is between 3 to 5 times bigger for MPI-ESM, BESM, CM2-SR2, CNRM-CM6, HadGEM3-LL, UKESM-LL and NorESM-MM. While in EC-Earth, it only accounts for 1/3 of the total model resolution (the remaining 2/3 are used for the ATM). Remarkably, the LM configuration used at NorESM uses a grid for the OCN which is 22 times bigger than the one for the ATM. As one would expect, the total number of grid points of an experiment can be explained solely by the ATM and OCN resolution used, but we will show later how adding more components/features (in yellow in Fig. 1) can have quite an impact on the performance anyway.

Fig. 2 shows the number of ocean and atmosphere grid points for the medium-high resolution experiments. We observe that like most of the low-resolution ones, all experiments use more grid points for the oceanic component than for the atmospheric one (notably, GFDL CM4-piC experiment use 55x more grid points in the OCN component). The ATM resolutions range between 1 and 0.4 degree, while OCN ones mostly run at 1/4 of a degree, except for the NERC-HadGEM3-GC3.1-HH experiment which runs the oceanic component at 1/12.

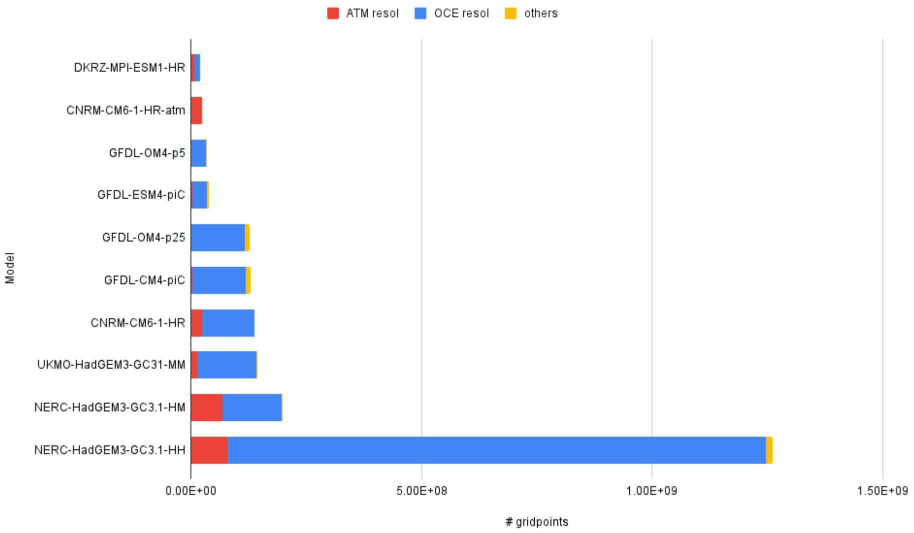

Figure 2. Atmosphere and ocean grid-points for medium-high resolution experiments

3.2 Complexity

From Tab. 5, we can also infer the *Complexity* (*Cmplx*) impact. The data reveals a wide variability in *Cmplx* across the models, with most models reporting a value that ranges between 100 to 400. Notably, GFDL (OM4 and CM4) and EC-Earth have considerably lower *Cmplx*. IPSL-CM6A model stands out in this context with a *Cmplx* of 750, which is markedly higher than the other models, potentially due to its representation of the carbon cycle. Likewise, we were expecting a much higher value for the EC-Earth-Veg experiment, but it was impossible to get this metric for the vegetation component (LPJ-Guess) even after contacting the developers. This highlights the challenge of obtaining this metric with accuracy, partly due to a lack of awareness of the number of prognostic variables of the components among users of the ESMs, leading to an overestimation for this metric, and also because the approximation based on the size of the restart files (Balaji et al., 2017, p. 25) is not always accurate. For instance, LPJ-Guess restart file size can measure tens of GB and depends on the Parallelization used for this component. What's

https://doi.org/10.5194/gmd-2023-188
Preprint. Discussion started: 23 October 2023

Table 5. Models that reported the Complexity metric

Institution	Model	Resol	SYPD	CHSY	Paral	Cmplx	Cpl C
GFDL	OM4-p25	1.26E+08	11.5	9748	4671	11	0.13
	OM4-p5	3.32E+07	15.9	1962	1300	13	0.14
	CM4	1.28E+08	9.98	15388	6399	31	0.26
	ESM4	3.76E+07	8.65	13576	4893	140	0.27
CNRM-CERFACS	CNRM-CM6-1-atm	2.98E+06	7.3	1292	393	128	
	CNRM-CM6-1	1.10E+07	8.1	1541	520	181	
	CNRM-CM6-1-HR-atm	2.36E+07	2.2	8520	781	128	
	CNRM-CM6-1-HR	1.37E+08	1.5	21552	1347	181	
	ESM2-1-atm	2.98E+06	7.1	1352	400	335	
	ESM2-1	1.10E+07	4.7	4289	840	393	
NERC	HadGEM3-GC3.1-HM	1.99E+08	0.58	192662	4656	54	0.21
	HadGEM3-GC3.1-HH	1.26E+09	0.49	588931	12024	54	
	HadGEM3-GC3.1-LL	1.14E+07	4.25	12198	2160	150	0.047
	UKESM1-AMIP	2.35E+06	1.64	7376	504	202	
	UKESM1-0-LL	1.14E+07	2.02	8554	720	252	0.078
UKMO	HadGEM3-GC31-LL	1.14E+07	4	13392	2232	228	0.061
	HadGEM3-GC31-MM	1.44E+08	1.65	62836	4320	236	0.105
	UKESM1-0-LL	1.14E+07	4.3	16074	2880	372	0.098
BSC	EC-Earth3	1.99E+07	15.2	1491	768	34	0.08
KNMI	EC-Earth3	1.99E+07	16.2	1286	868	34	
IPSL	IPSL-CM6A	1.06E+07	12	1900	950	750	0.05

more, explaining why NERC HadGEM3-GC31 *Cmplx* is almost 3 times larger for the lower resolution configuration (LL) than for the same experiment using more grid points (MM, HM and HH configurations) represents a challenge. Similarly, the notable differences between NERC and UMKO measurements, despite both running HadGEM-GC3.1 and UKESM1 models but on different platforms, raise questions about their source, which should require further investigation.

Nonetheless, the data from CNRM-CERFACS provides evidence supporting the idea that the *Cmplx* of a model should remain consistent regardless of the resolution, and only increase as additional features are simulated by the ESM. For instance, the *Cmplx* of CNRM-CM6 ATM standalone runs (CNRM-CM6-1-atm and CNRM-CM6-1-HR-atm) is 128 and grows up to 181 when the OCN component is included for the coupled configuration (CNRM-CM6-1 and CNRM-CM6-1-HR). The same is also observed for the CNRM-ESM2 model, where the *Cmplx* increases from 335 to 393 when adding the OCN component. Furthermore, in both cases, the ESMs require more PEs when running the coupled version. This shows a clear interconnection between the Parallelization and *Cmplx* as both will grow when comparing standalone and coupled simulations, other examples are: NERC standalone execution UKESM1-AMIP and UKESM1-LL coupled version, GFDL standalone OM4 (OCN only) runs and the coupled configurations ESM4 and CM4, and CNRM-CM6-atm (ATM only), CNRM-CM6-1 (ATM and OCN) and IPSL-CM6A (ATM, OCN and chemistry).

Therefore, *Cmplx* usually reduces the *SYPD* achieved and/or increases the *CHSY* given that adding a new component will, at best, only increase the latter. Maintaining the same throughput when increasing the *Cmplx* requires to use of more parallel resources, which translates into more costly executions and is usually correlated to parallel efficiency loss due to the need for coupling synchronizations and interpolations (e.g. see GFDL results in Tab. 5). The relation between *Cmplx* and the Coupling Cost is further discussed in subsection 3.5.

Table 6. Models that reported the metrics related to I/O

Institution	Model	Resol	Cmplx	SYPD	CHSY	Paral	DO	DI
BSC	EC-Earth3	1.99E+07	34	15.2	1213	768	1.12	0.0410
	EC-EarthVeg	1.99E+07		12.36	1491	768	1.13	0.0590
SMHI	EC-EarthVeg	1.99E+07		12.44	1667	864		0.0280
CMCC	CM2-SR5	6.94E+06	397	6.68	2069	576	1.04	0.050
GFDL	OM4-p5	3.32E+07	13	15.9	1962	1300		0.0392
	OM4-p25	1.26E+08	11	11.5	9748	4671		0.0178
	ESM4-piC	3.76E+07	140	8.65	13576	4893		0.0032
	CM4-piC	1.28E+08	31	9.98	15388	6399	1.24	0.0058
IMPE	IMPE-BESM	6.88E+06	132	3.6	1853	278		0.020
IPSL	IPSL-CM6A	1.06E+07	750	12	1900	950	1.2	0.070
NERC	HadGEM3-GC3.1-LL	1.14E+07	150	4.25	12198	2160	1.41	0.016
	HadGEM3-GC3.1-HM	1.99E+08	54	0.58	192662	4656		0.0006
	HadGEM3-GC3.1-HH	1.26E+09	54	0.49	588931	12024	1.41	0.0004
	UKESM1-AMIP	2.35E+06	202	1.64	7376	504	1.31	0.003
	UKESM1-0-LL	1.14E+07	252	2.02	8554	720	1.19	0.005
UKMO	UKESM1-0-LL	1.14E+07	372	4.3	16074	2880	1.03	0.019
	HadGEM3-GC31-LL	1.14E+07	228	4	13392	2232	1.03	0.074
	HadGEM3-GC31-MM	1.44E+08	236	1.65	62836	4320	1.02	0.050
NorESM	NorESM2-LM	1.01E+07		13.84	1665	960		0.065
	NorESM2-MM	1.14E+07		8.96	4886	1824		0.060

3.3 Data output

ESMs generate a large amount of output data, including model results, diagnostics, and intermediate variables, which need to be written to storage. Writing and saving this massive amount of data to disk or other storage mediums is time-consuming and can affect the overall performance of the model. Concurrent access to storage resources by multiple processes or multiple model instances can create contention, may represent an I/O bottleneck, and eventually degrade performance and scalability. CPMIP metrics add two metrics to quantify and evaluate the I/O workload: the *Data Output Cost* (*DO*), which reflects the cost of performing I/O and is determined as the ratio of *CHSY* with and without I/O; and the *Data Intensity* (*DI*), which measures the data production efficiency in terms of data generated per compute hour (i.e. GB/Core-hour).

Data Output Cost

From Tab. 6, we see that all the experiments conducted by UKMO and CMCC reported a *DO* below 1.05, even though the *DI* varies considerably between the different experiments. Moreover, we observe that the *DO* is much higher for the same ESM (HadGEM-GC31-LL and UKESM1-0-LL) when executed by NERC, reaching 1.19 for UKESM1-0-LL and 1.41 for HadGEM3-GC31-LL. It is not possible to know, however, if this is due to the difference between the HPC platform used or to differences in the model I/O configuration. This underscores the importance of of the specific model's I/O configuration in influencing the *DO* metirc. Besides, neither the metrics collected from UKMO nor the ones reported from NERC show that the *DO* should increase when running higher-resolution experiments (HadGEM3-GC31-MM and HH configurations). Moreover, EC-Earth and EC-Earth-Veg *DO* measurements are almost the same, suggesting that adding the Vegetation model to EC-Earth does not increase the cost of the IO, while UKESM runs conducted by NERC show that the *DO* is much higher when running the ATM standalone configuration, UKESM-AMIP, that the coupled run, UKESM-1-LL. Thus, the increase in Complexity or Resolution does not increase the cost of the IO but the cost of the whole ESM simulation, which can diminish the *DO* metric if IO workload stays constant.

Data Intensity

As seen in Tab. 6, the *DI* is generally of the order of MB per core-hour and gets smaller as we move to higher-resolution experiments (i.e. higher *CHSY*), meaning that the amount of data generated does not grow proportionally with the number of grid points nor with the execution cost. For instance, the *DI* reported for NERC-HadGEM, UKMO-HadGEM, NorESM2 and GFLD-OM4 experiments decreases when increasing the resolution. Thus, we observe a positive correlation between the *SYPD* and the *DI*.

3.4 Workflow and infrastructure costs

The real execution time of climate experiments can not be explained only by the speed at which a model can run. Queue times before having access to the HPC resources (usually managed by an external scheduler), service disruption, errors in the model/workflow manager, etc. can heavily extend the time-to-solution of ESMs. From the data in Tab. 1, we see that the difference between the *SYPD* and *ASYPD* reported varies a lot between institutions. Some claim that they had no overhead in their runs (KNMI), while for others it can account up to 78% (NorESM2-LR). The histogram in Fig. 3 helps illustrate the spread of the *ASYPD* overhead: it rarely surpasses 50% and half of the institutions reported it to be less than 20%. Judging from the spread of this metric and from the discussions after the collection, we consider that there are two groups: 1) Institutions that included solely the queue time, which reported an overhead under 20%, and 2) Institutions including not only the queue time but also the system interruptions and/or workflow management, which reported much higher values.

The results support the idea that queuing time represents an increment of around 10-20% to the speed of the ESM. On the other hand, adding interruptions and workflow management the total execution time could increase up to 40-50% compared to the simulation time alone. We do not have enough supporting data to draw any definitive conclusions, so we believe that it would be essential to add finer granularity to the *ASYPD* metric to be able to differentiate both factors. BSC CMIP6 results using the same configuration on two different platforms (Marenostrum and CCA) proved that the percentage of each part (queue time, interruptions or post-processing) could change among platforms even though the CMIP6 experiment is the same[3]. From the metrics listed in Tab. 3, we see that the difference between *SYPD* and *ASYPD* for the same model can significantly vary depending on the machine used for execution. For EC-Earth3 (standard and Vegetation experiments), the overhead ranges from 0% at KNMI to 0.35-0.40% at BSC and up to 0.47% at SMHI. However, it is important to note that the value provided

[3] https://shorturl.at/lzAHO, retrieved October 23, 2023

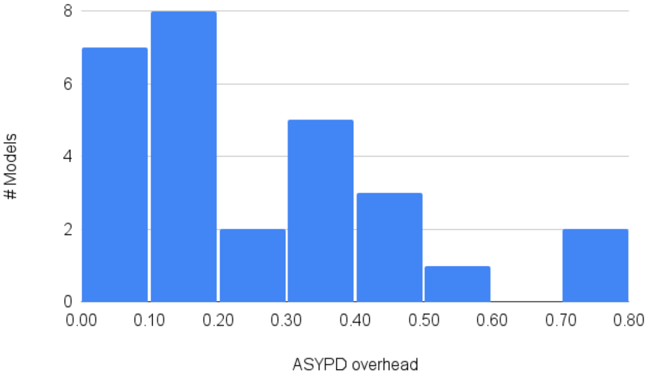

Figure 3. Histogram of the ASYPD overhead

by KNMI only accounts for the queue time, and they reported having instant access to the HPC resources. Furthermore, for HadGEM3-GC3.1-LL, we observe that NERC and UKMO runs are similar in the model execution speed, achieving approx. 4 *SYPD*, but totally different in the *ASYPD*. The overhead due to the workflow at UKMO is just 11%, whereas at NERC it takes 75%. We see something similar when comparing the same institutions for the UKESM-LL execution, where the overhead in UKMO is almost the same as before (16%), but it has drastically decreased at NERC. As we expected, the *ASYPD* overhead is related to the model *SYPD*, but more importantly to the workload of the platform used for the runs. Furthermore, we observed that for UKMO and MPI the smaller the parallelization, the smaller the overhead due to the workflow.

3.5 Coupling Cost

Coupling Cost (*Cpl C*) is an essential metric evaluated in this study. It quantifies the overhead introduced by coupling within an Earth System Model (ESM). This overhead encompasses various factors, including the coupling algorithms used for grid interpolations and calculations for conservative coupling. Additionally, it incorporates the impact of the load imbalance, which arises when different independent components of the ESM finish their computations at varying rates, potentially leaving PEs idle. The *Cpl C* metric is measured by comparing the time-processor integral for the entire model with the sum of all the individual concurrent components. Fig. 4 shows the list of institutions ordered from lower to higher *Cpl C*. Most institutions reported that the cost increase due to the coupling accounts for around 5-15% of the total. Only 4 (over the 16 that reported this metric) show an increase of over 20%. The data from GFDL (OM4-p5, OM4-p25, ESM4-piC, and CM4-piC) and UKMO (UKESM-LL and UKESM-AMIP) suggests that the increase in Complexity leads to higher *Cpl C* and lower *SYPD*. This aligns with the expectations, as the addition of a new component to the ESM will likely slow down the model and make the load balancing harder. It is noteworthy that a similar trend is observed in EC-Earth experiments. Even though we don't know the exact value for EC-EarthVeg *Cmplx*, it is known to be higher than in the standard EC-Earth (ATM-OCN) configuration due to the inclusion of Vegetation and Chemistry models. When comparing the performance of these two runs, we see a decrease in the *SYPD* and a concurrent increase in the *Cmplx* and *Cpl. C*, as discussed in more detail in subsection 3.2.

In general, the *Cpl C* tends to rise when running experiments that use a higher Parallelization. This could reflect a problem in the coupling phase. It can occur that the coupling algorithm is not scaling correctly or that the higher resolution configuration is not well-balanced. It is also likely that since the computing cost of running configurations in lower resolutions is smaller and less time-consuming, institutions can afford to run more spin-up tests and come up with a better distribution of processes among the coupled components to obtain a better load balance. In comparison, the contrary will happen for higher resolutions. Since there are no specific tools to balance a coupled model, these institutions are forced to use a trial-and-error approach, which is not trivial for complex configurations with several components and/or differences in the time-stepping among them.

For these cases, a finer granularity in the *Cpl C* metric and new ways to achieve a well-balanced configuration could be needed, splitting interpolation algorithm and waiting time in different sub-metrics or providing some of the CPMIPs (*SYPD*, *CHSY*...) not only for the coupled version but also per component.

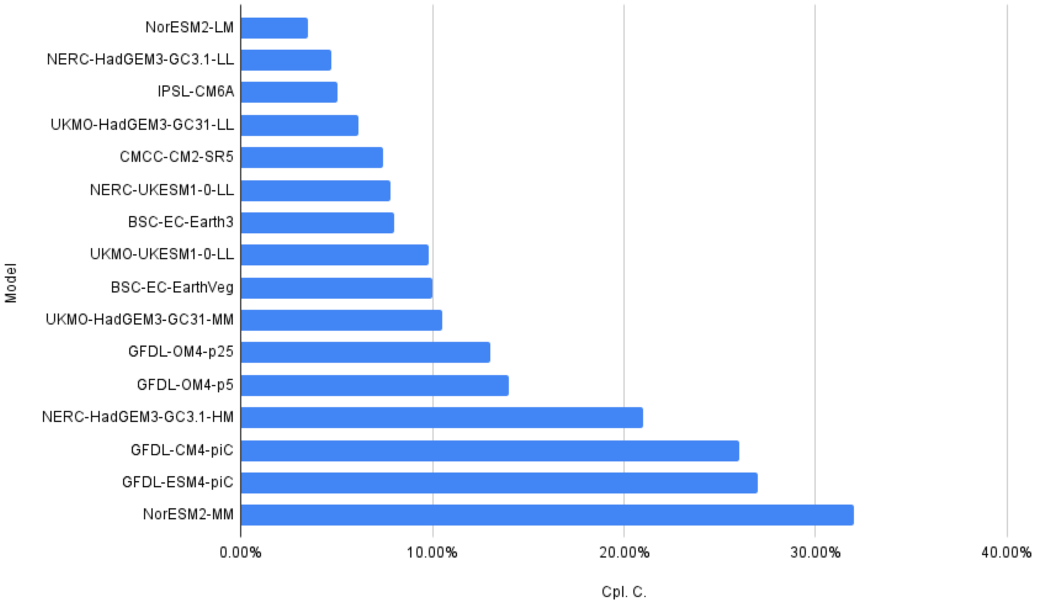

Figure 4. Coupling cost for all the institutions that provided the metric

3.6 Speed, cost, and parallelization

The *speed* of execution (*SYPD*) of a model is a fundamental metric that requires careful consideration. However, taken alone, it may not be enough to shed light on the model's performance itself. The meaning of a model's speed can only be fully understood when correlated to other important metrics. Among these the *parallelization* (*Paral*, i.e. the number of parallel

Table 7. Metrics for models available on different HPC platforms

Model	Institution	Resol	SYPD	CHSY	Paral
EC-Earth3	BSC	1.99E+07	15.2	1213	768
	KNMI	1.99E+07	16.2	1286	868
EC-Earth3Veg	BSC	1.99E+07	12.36	1491	768
	KNMI	1.99E+07	12.44	1667	864
UKESM1-0-LL	NERC	1.14E+07	2.02	8554	720
	UKMO	1.14E+07	4.3	16074	2880
HadGEM3-GC3.1-LL	NERC	1.14E+07	4.25	12198	2160
	UKMO	1.14E+07	4	13392	2232

resources allocated) stands out as a factor closely related to the speed and that, at the same time, directly influences the computational *cost* (*CHSY*) of the model execution. In this section, we show a detailed analysis of these three interconnected metrics. Contrary to what one would expect, the *SYPD* achieved by the models in this study is not always related to the resolution used nor to the *Paral* allocated. Although if we analyze how the same model performs on different HPC machines (Tab. 7), we note that higher values of *Paral* usually correspond to faster but more energy-consuming simulations.

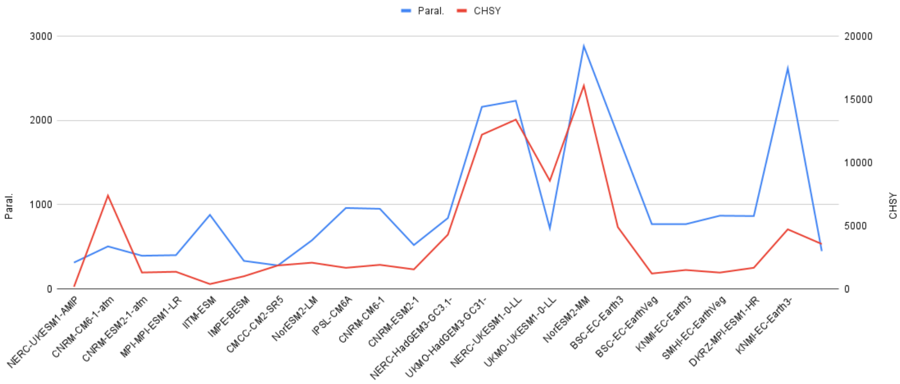

Figure 5. Low-resolution models Parallelization and CHSY

As seen in Fig. 5, the *Paral* and the *CHSY* are closely correlated in low-resolution models, showing that models do not scale in the current generation of HPC platforms. Otherwise, one would see that the CHSY of ESMs with similar *Resol* do not increase when using more processors given that the models run faster (i.e. higher *SYPD*). Also, as illustrated in Fig. 6, the level of parallelization tends to increase as we move to higher-resolution experiments. Thus, and given that we do not observe a relation between the Resolution and the *SYPD* achieved, we conclude that most institutions try to maintain at high-medium resolution the same *SYPD* achieved when running lower-resolution configurations, at the cost of increasing the *CHSY*.

In addition, and already discussed in subsection 3.5, the *Cpl C* grows together with the Parallelization, although there isn't any sign that it limits the speed of the models.

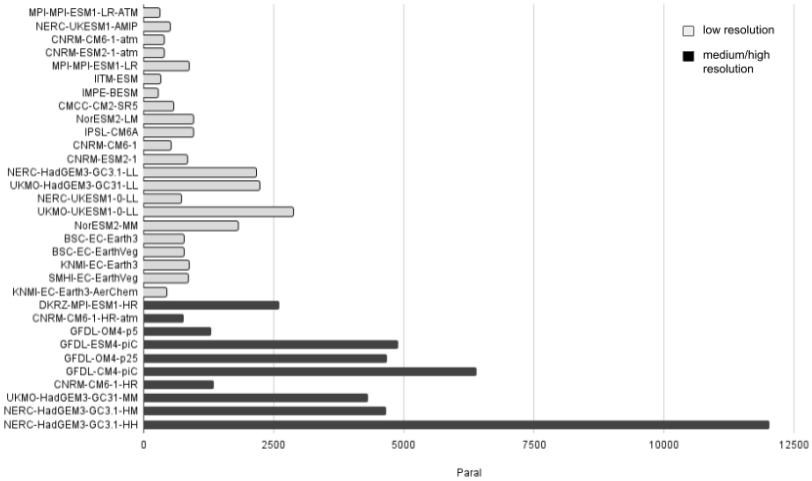

Figure 6. Parallelization for low (grey) and medium-high (black) resolution models

Table 8. Resolution, SYPD, CHSY, Paral and Memory Bloat results for UKESM, EC-Earth and HadGEM3-GC31 experiments

Model	Resol	SYPD	CHSY	Paral	Mem. Bloat
NERC-UKESM-AMIP	2.35E+06	1.6	7376	504	52.5
NERC-UKESM-LL	1.14E+07	2	8554	720	28
BSC-EC-Earth3	1.99E+07	15.2	1213	768	59.5
BSC-EC-EarthVeg	1.99E+07	12.4	1491	768	68.48
NERC-HadGEM3-GC3.1-LL	1.14E+07	4.3	12198	2160	56.8
UKMO-HadGEM3-GC31-LL	1.14E+07	4	13392	2232	46
UKMO-HadGEM3-GC31-MM	1.44E+08	1.7	62836	4320	120
NERC-HadGEM3-GC3.1-HM	1.99E+08	0.6	192662	4656	154
NERC-HadGEM3-GC3.1-HH	1.26E+09	0.5	588931	12024	183

3.7 Other CMIP6 measurements

In this section, we discuss the impact of *Memory Bloat* (Mem B) on model performance. The *Mem B* is defined as the ratio of actual memory size, obtained from the runtime memory usage (often referred “resident set size”, or RSS) to the ideal memory size, which represents the size of the complete model state. We show in Tab. 8 the *Mem B* values reported for some models. We observe how the *Mem B* increases with the resolution (e.g. NERC-HadGEM31) and when the *Cmplx* increases but the *Paral* remains the same (e.g. BSC-EC-Earth3 with and without the Vegetation model). Running in higher resolutions increases the *Mem B* likely due to the subdomain assigned to each compute unit getting bigger, and simulating more features implies keeping more data in memory. Additionally, it is worth noting that the *Cpl C* is directly linked to the Parallelization but not to the *CHSY* nor to the *SYPD*. This highlights the challenge represented by maintaining experiments load-balanced, especially when running on higher processor counts, and has been discussed in subsection 3.5.

https://doi.org/10.5194/gmd-2023-188
Preprint. Discussion started: 23 October 2023

Table 9. Other CMIP6 measurements

Institution	Useful Simulated Years*	Total Simulated Years	Useful Data Produced (PB)	Total Data Produced (PB)	Useful core hours (millions)	Total core hours (millions)	Total Person/Months	Total Energy Cost (TeraJoules)	PUE	CF (g CO2/kWh)	Carbon Footprint (tons CO2)
CMCC	965		0.097		1.99		7	1.61	1.84	408	329
CNRM-CERFACS	47,000		1.35	2.48	160	365	450	6.18	1.43	40	97
DKRZ	1,276	1,321	0.6		5.52	5.9		0.41	1.19	184	24
EC-Earth	28,105	38,854	0.8	1.41	31.13	46.36	115	1.24	1.35	357	165
IPSL	75,000	165,000	1.8	7.6	150	320	200	8.72	1.43	50	172
MPI-M	24,175	35,000	1.93		16.31			0.62	1.19	184	37
NCC-NorESM2	23,096		0.6		27.23	80	150	1.69			
NERC	640		0.46		55.5			2.17	1.10	0	0
UKMO	37,237		10.4		683			26.7	1.35	87	868

. *The Useful Simulated Years column values can differ from Tab. 1 given that some of the experiment runs were not shown in that table

3.8 Carbon footprint

In addition to the CPMIP collection, we have also gathered the general metrics shown in Tab. 10. These metrics provide both useful (only accounting for simulations that produced data with scientific value) and total (encompassing all simulations, including spin-up and any runs that were finally discarded) numbers for the complete execution of CMIP6 experiments at the different institutions. They can be used to provide an idea about the total and useful number of years simulated, data produced and core hours consumed to finish the European community CMIP6 experiments. Although we did our best to collect the most updated data, we are aware that these numbers could have changed since the data collection was finished. We know that some institutions were doing some minor and final executions and updating databases such as ESGF. However, we consider Tab. 10 a very good representation of the effort done for the collection during CMIP6. In any case and taking into account the previous reasons, we do not analyze the results themselves and we will use this information to evaluate the Carbon Footprint associated with running models for large-scale projects like CMIP6, which is also a very interesting example for the community. By considering the useful Simulated Years, the HPC machine efficiency, and the KWH to CO2 conversion rates provided by each energy supplier, we calculated the Carbon Footprint (in tons of CO2) using Eq. 1. As the reader can see, NERC reported a zero Carbon Footprint due to their green tariff power supplier. Among other institutions, CMCC is the one with the highest CF, followed by EC-Earth. Both significantly surpass the emissions of the other institutions: CERFACS, MPI, and UKMO have very small CO2 emissions per kWh. Regarding machine efficiency, CMCC reported that Zeus is the least power-efficient machine, with a Power Usage Effectiveness (PUE) of 1.84. CERFACS, IPSL, EC-Earth and UKMO reported similar values for their machines, while DKRZ, MPI-M and NERC have reported a PUE under 1.2. We believe that CMCC's Carbon Footprint may be overestimated, considering they simulated fewer than 1000 years yet reported nearly double the CO2 emissions compared to EC-Earth or IPSL, despite these institutions having simulated longer experiments (in SY). The Total Energy Cost of UKMO seems too big compared to their reported Useful SY. However, this can be attributed to the cost of maintaining the useful data produced, which amounts to 10.4 PB of disk space. The total Carbon Footprint is 1692 tCO2, representing data from 8 out of 45 institutions. This is equivalent to driving 377 gasoline cars non-stop for a year or consuming 720,705 L of gasoline at once.

https://doi.org/10.5194/gmd-2023-188
Preprint. Discussion started: 23 October 2023

4 Drawbacks and actions recommended

Thanks to the experience learned from the data collection and analysis done, we recognize the importance of highlighting the specific drawbacks we have found during this first collection as well as our recommendations to improve the collection and analysis for future iterations of multi-model climate research projects, such as CMIP7. The authors will continue working on this topic in the future not only to provide new approaches to facilitate the collection, but also in fostering the collaboration of the weather and climate science community to address the computational challenges of Earth modelling.

Table 11. Drawbacks and Recommended Actions for CMIP6 Metrics

Drawbacks	Recommended Actions
CPMIPs are not enough to compare the performance of different ESMs running on different HPC platforms.	Multi-model comparisons will be better grounded once more data is available. Integrating the CPMIPs in dwarfs like the High Performance Climate and Weather (HPCW, van Werkhoven et al., 2023) benchmark to evaluate the performance of the different machines used by the community.
Lack of resources and time to collect metrics after CMIP experiments.	Perform metric collection before or during CMIP experiments. Develop portable and automated processes for efficient collection.
Inconsistencies in metric collection hinder inter-model comparisons.	Normalize metric collection methods across institutions before multi-model runs. Develop tools for automatize the collection (e.g. integrated into the workflow manager).
Difficulty in identifying computational bottlenecks due to limited information.	Split sensitive metrics into sub-metrics for finer analysis. For instance, the Coupling Cost should separate interpolation from load-imbalance cost, and the ASYPD should differentiate between queue time and system interruptions.

5 Conclusions

One of the limiting factors of climate science is the computational performance that Earth System Models (ESM) can achieve in the current HPC platforms. This limitation imposes constraints on the number of years that can be simulated, the number of ensembles used, the resolution used by the models, the number of features simulated in one experiment, IO intensity, data diagnostics calculated during the run, etc. Evaluating the performance of an ESM is a tremendous amount of work that generally requires: profiling the application, using tools to visualize and understand the profiling information, and developing and applying solutions based on the bottlenecks found. This process becomes even more complex when dealing with models used in large-scale multi-model projects like CMIP6, where multiple ESM are executed by different institutions that have access to diverse HPC platforms. To address these challenges, the Computational Performance Model Intercomparison Project

(CPMIP) metrics were designed to be: universally available, easy to collect, representative of the actual performance of ESMs and of the entire life-cycle of modeling (i.e. simulation and workflow costs).

This paper presents, for the first time, the results obtained from the CPMIP collection during the CMIP6 exercise, providing the list of partners involved, ESM configurations, as well as the CPMIP metrics per institution/model. Furthermore, it goes well beyond mere data presentation and offers in-depth analysis for each metric collected to demonstrate the broader utility of the CPMIP collection. For instance, the study investigates the resolution used by each model on the OCN and ATM components, explores the relationship between execution speed and cost with the other metrics, assesses the impact of running models with higher parallelization, complexity, or I/O requirements, examines the overhead caused by queuing and workflow management, explores the coupling cost across different configurations, etc.

Besides the CPMIP metrics, this paper highlights results from collaborations with other groups, such as the Carbon Footprint. These collaborations underscore the shared concern of multiple institutions regarding the computational performance in climate science and the joint effort to estimate the carbon footprint of the simulations conducted during the CMIP6 exercise.

Finally, the paper addresses the main issues and drawbacks encountered during the collection and analysis of the metrics. These points should be of particular interest to the partners, aiming to improve and facilitate future collections. The paper also proposes recommendations to confront these challenges, which can be adopted by the community for the development of novel tools and more finely-grained metrics which would facilitate upcoming similar works. Continuous collection of these metrics in future multi-model projects will facilitate the development of a shared database for the community and vendors.

Data availability. The original data used during this work can be found here: http://bit.ly/3Y6XhHM

Author contributions. M.C. Acosta led the data collection process, ensuring that multiple institutions provided the necessary data and providing technical assistance throughout. S. Palomas conducted the data analysis and validation, and took the lead in writing the manuscript. S.P. Ticco contributed to data analysis and played a key role in revising and correcting the manuscript. Other co-authors were responsible for data collection from their respective models and institutions

Competing interests. The contact author declares that Sophie Valcke, a co-author of this article, serves as a member of the editorial board of the Geoscientific Model Development (GMD) journal. The authors have no other competing interests to declare.

Acknowledgements. The research leading to these results has received funding from the EU H2020 ISENES3, under grant agreement n° 824084 and co-funding from the National Research Agency through OEMES (PID2020-116324RA-I00)

https://doi.org/10.5194/gmd-2023-188
Preprint. Discussion started: 23 October 2023

References

Balaji, V., Maisonnave, E., Zadeh, N., Lawrence, B. N., Biercamp, J., Fladrich, U., Aloisio, G., Benson, R., Caubel, A., Durachta, J., Foujols, M.-A., Lister, G., Mocavero, S., Underwood, S., and Wright, G.: CPMIP: measurements of real computational performance of Earth system models in CMIP6, Geoscientific Model Development, 10, 19–34, https://doi.org/10.5194/gmd-10-19-2017, 2017.

Bauer, P., Dueben, P. D., Hoefler, T., Quintino, T., Schulthess, T. C., and Wedi, N. P.: The digital revolution of Earth-system science, Nature Computational Science, 1, 104–113, https://doi.org/10.1038/s43588-021-00023-0, 2021.

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Caubel, A., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., D'Andrea, F., Davini, P., de Lavergne, C., Denvil, S., Deshayes, J., Devilliers, M., Ducharne, A., Dufresne, J.-L., Dupont, E., Éthé, C., Fairhead, L., Falletti, L., Flavoni, S., Foujols, M.-A., Gardoll, S., Gastineau, G., Ghattas, J., Grandpeix, J.-Y., Guenet, B., Guez, Lionel, E., Guilyardi, E., Guimberteau, M., Hauglustaine, D., Hourdin, F., Idelkadi, A., Joussaume, S., Kageyama, M., Khodri, M., Krinner, G., Lebas, N., Levavasseur, G., Lévy, C., Li, L., Lott, F., Lurton, T., Luyssaert, S., Madec, G., Madeleine, J.-B., Maignan, F., Marchand, M., Marti, O., Mellul, L., Meurdesoif, Y., Mignot, J., Musat, I., Ottlé, C., Peylin, P., Planton, Y., Polcher, J., Rio, C., Rochetin, N., Rousset, C., Sepulchre, P., Sima, A., Swingedouw, D., Thiéblemont, R., Traore, A. K., Vancoppenolle, M., Vial, J., Vialard, J., Viovy, N., and Vuichard, N.: Presentation and Evaluation of the IPSL-CM6A-LR Climate Model, Journal of Advances in Modeling Earth Systems, 12, e2019MS002 010, https://doi.org/https://doi.org/10.1029/2019MS002010, 2020.

Dennis, J. M., Vertenstein, M., Worley, P. H., Mirin, A. A., Craig, A. P., Jacob, R., and Mickelson, S.: Computational performance of ultra-high-resolution capability in the Community Earth System Model, The International Journal of High Performance Computing Applications, 26, 5–16, 2012.

Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arsouze, T., Bergman, T., Bernardello, R., Boussetta, S., Caron, L.-P., Carver, G., Castrillo, M., Catalano, F., Cvijanovic, I., Davini, P., Dekker, E., Doblas-Reyes, F. J., Docquier, D., Echevarria, P., Fladrich, U., Fuentes-Franco, R., Gröger, M., v. Hardenberg, J., Hieronymus, J., Karami, M. P., Keskinen, J.-P., Koenigk, T., Makkonen, R., Massonnet, F., Ménégoz, M., Miller, P. A., Moreno-Chamarro, E., Nieradzik, L., van Noije, T., Nolan, P., O'Donnell, D., Ollinaho, P., van den Oord, G., Ortega, P., Prims, O. T., Ramos, A., Reerink, T., Rousset, C., Ruprich-Robert, Y., Le Sager, P., Schmith, T., Schrödner, R., Serva, F., Sicardi, V., Sloth Madsen, M., Smith, B., Tian, T., Tourigny, E., Uotila, P., Vancoppenolle, M., Wang, S., Wårlind, D., Willén, U., Wyser, K., Yang, S., Yepes-Arbós, X., and Zhang, Q.: The EC-Earth3 Earth system model for the Coupled Model Intercomparison Project 6, Geoscientific Model Development, 15, 2973–3020, https://doi.org/10.5194/gmd-15-2973-2022, 2022.

Du, M., Zheng, F., Zhu, J., Lin, R., Yang, H., and Chen, Q.: A New Ensemble-Based Approach to Correct the Systematic Ocean Temperature Bias of CAS-ESM-C to Improve Its Simulation and Data Assimilation Abilities, Journal of Geophysical Research: Oceans, 125, e2020JC016 406, https://doi.org/https://doi.org/10.1029/2020JC016406, e2020JC016406 2020JC016406, 2020.

Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., Krasting, J. P., Malyshev, S., Naik, V., Paulot, F., Shevliakova, E., Stock, C. A., Zadeh, N., Balaji, V., Blanton, C., Dunne, K. A., Dupuis, C., Durachta, J., Dussin, R., Gauthier, P. P. G., Griffies, S. M., Guo, H., Hallberg, R. W., Harrison, M., He, J., Hurlin, W., McHugh, C., Menzel, R., Milly, P. C. D., Nikonov, S., Paynter, D. J., Ploshay, J., Radhakrishnan, A., Rand, K., Reichl, B. G., Robinson, T., Schwarzkopf, D. M., Sentman, L. T., Underwood, S., Vahlenkamp, H., Winton, M., Wittenberg, A. T., Wyman, B., Zeng, Y., and Zhao, M.: The GFDL Earth System Model Version 4.1 (GFDL-ESM 4.1): Overall Coupled Model Description and Simulation Characteristics, Journal of Advances in Modeling Earth Systems, 12, e2019MS002 015, https://doi.org/https://doi.org/10.1029/2019MS002015, 2020.

Evans, J. P., Ji, F., Abramowitz, G., and Ekström, M.: Optimally choosing small ensemble members to produce robust climate simulations, Environmental Research Letters, 8, 044 050, 2013.

Evensen, G.: The ensemble Kalman filter: Theoretical formulation and practical implementation, Ocean dynamics, 53, 343–367, 2003.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Freedman, D., Pisani, R., and Purves, R.: Statistics (international student edition), Pisani, R. Purves, 4th edn. WW Norton & Company, New York, 2007.

Fuhrer, O., Osuna, C., Lapillonne, X., Gysi, T., Cumming, B., Bianco, M., Arteaga, A., and Schulthess, T. C.: Towards a performance portable, architecture agnostic implementation strategy for weather and climate models, Supercomputing Frontiers and Innovations, 1, 45–62, https://doi.org/10.14529/jsfi140103, 2014.

Giles, M. and Reguly, I.: Trends in high-performance computing for engineering calculations, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 372, 20130 319, 2014.

Hill, C., DeLuca, C., Balaji, Suarez, M., and Da Silva, A.: The architecture of the Earth System Modeling Framework, Computing in Science & Engineering, 6, 18–28, https://doi.org/10.1109/MCISE.2004.1255817, 2004.

Joussaume, S.: IS-ENES: Infrastructure for the European Network for Earth System Modelling, in: EGU General Assembly Conference Abstracts, EGU General Assembly Conference Abstracts, p. 6039, 2010.

Krishnan, R., Swapna, P., Choudhury, A. D., Narayansetti, S., Prajeesh, A., Singh, M., Modi, A., Mathew, R., Vellore, R., Jyoti, J., et al.: The IITM earth system model (IITM ESM), arXiv preprint arXiv:2101.03410, 2021.

Lovato, T., Peano, D., Butenschön, M., Materia, S., Iovino, D., Scoccimarro, E., Fogli, P. G., Cherchi, A., Bellucci, A., Gualdi, S., Masina, S., and Navarra, A.: CMIP6 Simulations With the CMCC Earth System Model (CMCC-ESM2), Journal of Advances in Modeling Earth Systems, 14, e2021MS002 814, https://doi.org/https://doi.org/10.1029/2021MS002814, 2022.

McGuffie, K. and Henderson-Sellers, A.: Forty years of numerical climate modelling, International Journal of Climatology, 21, 1067–1109, https://doi.org/https://doi.org/10.1002/joc.632, 2001.

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., Bunzel, F., Esch, M., Ghosh, R., Haak, H., Ilyina, T., Kleine, T., Kornblueh, L., Li, H., Modali, K., Notz, D., Pohlmann, H., Roeckner, E., Stemmler, I., Tian, F., and Marotzke, J.: A Higher-resolution Version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR), Journal of Advances in Modeling Earth Systems, 10, 1383–1413, https://doi.org/https://doi.org/10.1029/2017MS001217, 2018.

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., Debernard, J. B., Gupta, A. K., He, Y.-C., Kirkevåg, A., Schwinger, J., Tjiputra, J., Aas, K. S., Bethke, I., Fan, Y., Griesfeller, J., Grini, A., Guo, C., Ilicak, M., Karset, I. H. H., Landgren, O., Liakka, J., Moseid, K. O., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iversen, T., and Schulz, M.: Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations, Geoscientific Model Development, 13, 6165–6200, https://doi.org/10.5194/gmd-13-6165-2020, 2020.

Sellar, A. A., Walton, J., Jones, C. G., Wood, R., Abraham, N. L., Andrejczuk, M., Andrews, M. B., Andrews, T., Archibald, A. T., de Mora, L., Dyson, H., Elkington, M., Ellis, R., Florek, P., Good, P., Gohar, L., Haddad, S., Hardiman, S. C., Hogan, E., Iwi, A., Jones, C. D., Johnson, B., Kelley, D. I., Kettleborough, J., Knight, J. R., Köhler, M. O., Kuhlbrodt, T., Liddicoat, S., Linova-Pavlova, I., Mizielinski, M. S., Morgenstern, O., Mulcahy, J., Neininger, E., O'Connor, F. M., Petrie, R., Ridley, J., Rioual, J.-C., Roberts, M., Robertson, E., Rumbold, S., Seddon, J., Shepherd, H., Shim, S., Stephens, A., Teixiera, J. C., Tang, Y., Williams, J., Wiltshire, A., and Griffiths,

P. T.: Implementation of U.K. Earth System Models for CMIP6, Journal of Advances in Modeling Earth Systems, 12, e2019MS001 946, https://doi.org/https://doi.org/10.1029/2019MS001946, 2020.

Séférian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., Decharme, B., Delire, C., Berthet, S., Chevallier, M., Sénési, S., Franchisteguy, L., Vial, J., Mallet, M., Joetzjer, E., Geoffroy, O., Guérémy, J.-F., Moine, M.-P., Msadek, R., Ribes, A., Rocher, M., Roehrig, R., Salas-y Mélia, D., Sanchez, E., Terray, L., Valcke, S., Waldman, R., Aumont, O., Bopp, L., Deshayes, J., Éthé, C., and Madec, G.: Evaluation of CNRM Earth System Model, CNRM-ESM2-1: Role of Earth System Processes in Present-Day and Future Climate, Journal of Advances in Modeling Earth Systems, 11, 4182–4227, https://doi.org/https://doi.org/10.1029/2019MS001791, 2019.

Takizawa, H., Egawa, R., Takahashi, D., and Suda, R.: HPC refactoring with hierarchical abstractions to help software evolution, in: Sustained Simulation Performance 2012: Proceedings of the joint Workshop on High Performance Computing on Vector Systems, Stuttgart (HLRS), and Workshop on Sustained Simulation Performance, Tohoku University, 2012, pp. 27–33, Springer, 2013.

van Werkhoven, B., van den Oord, G., Sclocco, A., Heldens, S., Azizi, V., Raffin, E., Guibert, D., Lucido, L., Moulard, G.-E., Giuliani, G., van Stratum, B., and van Heerwaarden, C.: To make Europe's Earth system models fit for exascale - Deliverable D3.5, https://doi.org/10.5281/zenodo.7671032, ESiWACE2 stands for Centre of Excellence in Simulation of Weather and Climate in Europe Phase 2. ESiWACE2 is funded by the European Union's Horizon 2020 research and innovation programme (H2020-INFRAEDI-2018-1 call) under grant agreement 823988., 2023.

Veiga, S. F., Nobre, P., Giarolla, E., Capistrano, V., Baptista Jr., M., Marquez, A. L., Figueroa, S. N., Bonatti, J. P., Kubota, P., and Nobre, C. A.: The Brazilian Earth System Model ocean–atmosphere (BESM-OA) version 2.5: evaluation of its CMIP5 historical simulation, Geoscientific Model Development, 12, 1613–1642, https://doi.org/10.5194/gmd-12-1613-2019, 2019.

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., Colin, J., Guérémy, J.-F., Michou, M., Moine, M.-P., Nabat, P., Roehrig, R., Salas y Mélia, D., Séférian, R., Valcke, S., Beau, I., Belamari, S., Berthet, S., Cassou, C., Cattiaux, J., Deshayes, J., Douville, H., Ethé, C., Franchistéguy, L., Geoffroy, O., Lévy, C., Madec, G., Meurdesoif, Y., Msadek, R., Ribes, A., Sanchez-Gomez, E., Terray, L., and Waldman, R.: Evaluation of CMIP6 DECK Experiments With CNRM-CM6-1, Journal of Advances in Modeling Earth Systems, 11, 2177–2213, https://doi.org/https://doi.org/10.1029/2019MS001683, 2019.

Wang, D. and Yuan, F.: High-Performance Computing for Earth System Modeling, pp. 175–184, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-47998-5_10, 2020.

Wang, D., Post, W., and Wilson, B.: Climate change modeling: Computational opportunities and challenges, Computing in Science & Engineering, 13, 36–42, 2010.

Williams, K. D., Copsey, D., Blockley, E. W., Bodas-Salcedo, A., Calvert, D., Comer, R., Davis, P., Graham, T., Hewitt, H. T., Hill, R., Hyder, P., Ineson, S., Johns, T. C., Keen, A. B., Lee, R. W., Megann, A., Milton, S. F., Rae, J. G. L., Roberts, M. J., Scaife, A. A., Schiemann, R., Storkey, D., Thorpe, L., Watterson, I. G., Walters, D. N., West, A., Wood, R. A., Woollings, T., and Xavier, P. K.: The Met Office Global Coupled Model 3.0 and 3.1 (GC3.0 and GC3.1) Configurations, Journal of Advances in Modeling Earth Systems, 10, 357–380, https://doi.org/https://doi.org/10.1002/2017MS001115, 2018.