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

Acosta, M. C. ORCID: https://orcid.org/0000-0001-7054-8168, Palomas, S. ORCID: https://orcid.org/0000-0002-2191-152X, Paronuzzi Ticco, S. V., Utrera, G., Biercamp, J. ORCID: https://orcid.org/0000-0002-6677-2164, Bretonniere, P.-A. ORCID: https://orcid.org/0000-0002-3066-6685, Budich, R. ORCID: https://orcid.org/0000-0002-9274-4052, Castrillo, M. ORCID: https://orcid.org/0000-0003-1826-623X, Caubel, A., Doblas-Reyes, F. ORCID: https://orcid.org/0000-0002-6622-4280, Epicoco, I., Fladrich, U., Joussaume, S., Kumar Gupta, A., Lawrence, B. ORCID: https://orcid.org/0000-0001-9262-7860, Le Sager, P., Lister, G., Moine, M.-P., Rioual, J.-C., Valcke, S. ORCID: https://orcid.org/0000-0002-0438-5978, Zadeh, N. and Balaji, V. ORCID: https://orcid.org/0000-0001-7561-5438 (2024) The computational and energy cost of simulation and storage for climate science: lessons from CMIP6. Geoscientific Model Development, 17 (8). pp. 3081-3098. ISSN 1991-9603 doi: https://doi.org/10.5194/gmd-17-3081-2024 Available at https://centaur.reading.ac.uk/116098/

Geosci. Model Dev., 17, 3081–3098, 2024
https://doi.org/10.5194/gmd-17-3081-2024

Geoscientific Model Development
Open Access
EGU

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

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

[1]Barcelona Supercomputing Center, Plaça d'Eusebi Güell, 1–3, 08034 Barcelona, Spain
[2]German Climate Computing Centre, Bundesstraße 45a, 20146 Hamburg, Germany
[3]Max Planck Institute, Hofgartenstr. 8, 80539 Munich, Germany
[4]Institut Pierre-Simon Laplace, 11 Bd d'Alembert, 78280 Guyancourt, France
[5]Euro-Mediterranean Center on Climate Change, Via della Libertà, 12, 30121 Venice, Italy
[6]Swedish Meteorological and Hydrological Institute, 601 76 Norrköping, Sweden
[7]Norwegian Research Centre, Nygårdsgaten 112, 5008 Bergen, Norway
[8]National Centre for Atmospheric Science, Fairbairn House, 71–75 Clarendon Rd, Woodhouse, Leeds LS2 9PH, United Kingdom
[9]Royal Netherlands Meteorological Institute, Utrechtseweg 297, 3731 GA De Bilt, the Netherlands
[10]European Center for Advanced Research and Training in Scientific Computing, 42 Av. Gaspard Coriolis, 31100 Toulouse, France
[11]Meteorological Office, Fitzroy Road, Exeter, Devon, EX1 3PB, United Kingdom
[12]National Oceanic and Atmospheric Administration, 1401 Constitution Avenue NW, Room 5128, Washington, DC 20230, USA
[13]High Meadows Environmental Institute, Princeton University, Guyot Hall, Room 129, Princeton, NJ 08544-1003, USA
[14]Catalan Institution for Research and Advanced Studies, Passeig Lluís Companys 23, 08010 Barcelona, Spain

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

Received: 6 October 2023 – Discussion started: 23 October 2023
Revised: 6 February 2024 – Accepted: 16 February 2024 – Published: 19 April 2024

Abstract. The Coupled Model Intercomparison Project (CMIP) is one of the biggest international efforts aimed at better understanding the past, present, and future of climate changes in a multi-model context. A total of 21 model intercomparison projects (MIPs) were endorsed in its sixth phase (CMIP6), which included 190 different experiments that were used to simulate 40 000 years and produced around 40 PB of data in total. This paper presents the main findings obtained from the CPMIP (the Computational Performance Model Intercomparison Project), a collection of a common set of metrics, specifically designed for assessing climate model performance. These metrics were exclusively collected from the production runs of experiments used in CMIP6 and primarily from institutions within the IS-ENES3 consortium. The document presents the full set of CPMIP metrics per institution and experiment, including a detailed analysis and discussion of each of the measurements. During the analysis, we found a positive correlation between the core hours needed, the complexity of the models, and the resolution used. Likewise, we show that between 5 %–15 % of the execution cost is spent in the coupling between independent components, and it only gets worse by increasing the number of resources. From the data, it is clear that queue times have a great impact on the actual speed achieved and have a huge variability across different institutions, ranging from none to up to 78 % execution overhead. Furthermore, our evaluation

shows that the estimated carbon footprint of running such big simulations within the IS-ENES3 consortium is 1692 t of CO_2 equivalent.

As a result of the collection, we contribute to the creation of a comprehensive database for future community reference, establishing a benchmark for evaluation and facilitating the multi-model, multi-platform comparisons crucial for understanding climate modelling performance. Given the diverse range of applications, configurations, and hardware utilised, further work is required for the standardisation and formulation of general rules. The paper concludes with recommendations for future exercises aimed at addressing the encountered challenges which will facilitate more collections of a similar nature.

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 high-performance computing (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).

Virtually all models are designed to exploit the parallelism of HPC machines so that the results can be obtained 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^{-1}, memory bandwidth, I/O speed, etc.) climate software evolves much more slowly. Balaji (2015) and Liu et al. (2013) show how challenging it is to adapt multi-scale, multi-physics climate models to new hardware or programming paradigms. These models, often community-developed software, are very complex, inherently chaotic, and subject to numerical stability, all of which contribute to a slower evolution of the codes. Bauer et al. (2021) illustrate how climate science did take advantage of Moore's law (Bondyopadhyay, 1998) and Dennard scaling (Frank et al., 2001) without much pressure to fundamentally revise numerical methods and programming paradigms, leading to huge legacy codes mostly driven by scientific concerns. Consequently, such codes achieve notably poor sustained floating-point performance in present-day CPU architectures. Enhancing the performance of these models is crucial to boost the rate at which they can grow (in the resolution, complexity, and features represented). In a context where energy costs are rising, running faster and more cost-effective simulations is key to contributing to the advancement of climate research.

The performance of ESMs is hardly limited by only 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 are 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; and models that include more individual components will be limited by the load balance achieved between them and by the coupler performance.

Balaji et al. (2017) proposed a set of 12 performance metrics that define the Computation Performance for Model Intercomparison Project (CPMIP), designed explicitly for climate science by considering the structure of ESMs and how they are executed in real experiments. This set of metrics includes the climate experiment and platform properties; the computational speed and cost (core hours and energy); and measures for the coupling, I/O overhead, and memory consumption. Each one is described in detail in Table 1 and Sect. 3.

In this paper, we present in Sect. 2 the collection of CPMIP metrics from 33 experiments used for climate projections in the Coupled Model Intercomparison Project phase 6 (CMIP6). The collection effort has been predominantly led by institutions affiliated with the IS-ENES3 (Joussaume, 2010), a consortium founded by a Horizon 2020 (EU funding programme: https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en, last access: 2 February 2024) project composed of the most important weather and climate centres in Europe and devoted to improving the infrastructure to make the Earth System Grid Federation (ESGF) and CMIP publication easier. This compilation is the first of its kind and constitutes a representative part of the total 124 CMIP6 experiments, involving 45 institutions (http://esgf-ui.cmcc.it/esgf-dashboard-ui/data-archiveCMIP6.html, last access: 2 February 2024). Our data encompass 33 different experiments that were used to simulate almost 500 000 years during CMIP6 on 14 different HPC machines and involving 14 independent modelling institutions. All experiments are listed in Table 2, along with the institution in charge, the experiment name, HPC platform, ocean and atmosphere resolutions, and the main reference to the experiment configuration. In addition, Table 3 shows the complete collection of CPMIP metrics for each one of the models, and Table 4 lists the HPC machines that have been used to run these models. Furthermore, in Sect. 3, we include the analysis of the metrics to underscore the most significant insights derived from this data collection. We study in detail the measurements reported

by each institution, grouping them based on experiment configurations, establishing relationships between intertwined metrics, and discussing the strengths and difficulties encountered during the analysis of each metric. For instance, our analysis reveals that institutions tend to increase the number of resources used in higher-resolution experiments, thereby mitigating the expected increase in execution time at the expense of increasing the core hours required. Similarly, the addition of extra components simulated increases the core hours needed and the cost of coupling interactions and synchronisations between models as well. Institutions reported that the coupling cost entails an execution cost overhead typically ranging between 5 %–15 %, and it tends to be more problematic higher processor counts. Additionally, the numbers indicate that the volume of data generated by an experiment does not correspond to increases in resolution or core hours needed, contrary to expectations. We observed very different queue times for HPC resources across institutions, ranging from instantaneous access to introducing an execution time overhead of up to 78 %. Furthermore, we present an initial approximation of the carbon footprint generated from executing these experiments, totalling 1692 t of CO_2 equivalent.

Our study emphasises the significance of developing standardised metrics for assessing climate model performance. This contribution will serve to establish a database for future reference and that multiple institution modellers will be able to use for comparison, which we believe to be essential for evaluating climate modelling performance. The noise and variability present in the dataset are the results of the diversity of the applications represented and the hardware under study, making it challenging to formulate general rules. Despite these difficulties, our paper concludes with recommendations for future exercises aimed at addressing these challenges.

2 Data collection

The collection process was coordinated and supervised to get the metric results, including meetings, reporting, and surveys conducted at different stages of the CMIP6 simulations (before, during, and after the simulation runs). All the partners listed in Table 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 we provided support to those institutions that required it during the collection process.

The data collection was divided into two steps: the initial phase 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 (ES-DOC, https://es-doc.org/, last access: March 2024) in the middle of 2021, publishing CPMIP along with the other CMIP6 results.

As the reader can see, not all institutions managed to provide the full set of CPMIP performance metrics. The metrics frequently missing are the *coupling cost*, *memory bloat*, and *data output cost*. This is primarily attributed to the challenges involved in their collection compared to metrics like *SYPD* or parallelisation, which are well known within the community and relatively easier to obtain. Other impediments to collect the CPMIP metrics include time and resource constraints, particularly considering that the focus of the simulations leans more towards science aspects than towards the computational realm during CMIP6 runs. Additionally, some institutions reported that changes in the underlying computational infrastructure have made the collection process more difficult.

2.1 Additional data collected

The CPMIP metrics not only serve 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 consortium, which was responsible for evaluating the total energy cost associated with the CMIP6 experiments.

$$\text{Total energy cost} = \text{useful simulated years} \times \text{JPSY} \quad (1)$$

The total energy cost of an experiment is defined as the product of the useful simulated years, defined as years of simulation that produced data with a scientific value that was either shared between the groups or kept within the producer group for scientific analysis, and the joules per simulated year (JPSY). This collaboration enabled us to provide for the first time an estimation of the carbon footprint related to those experiments. The carbon footprint was calculated following Eq. (2).

$$\text{Carbon footprint} = \text{total energy cost} \times \text{CF} \times \text{PUE}, \quad (2)$$

where the total energy cost is in megawatt-hours; CF is the greenhouse gas conversion factor from megawatt-hours to kilograms of CO_2 according to the supplier bill or the country energy mix; and PUE (power usage effectiveness) accounts for other costs sustained from the data centre, 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 Table 9.

2.2 Uncertainty in the measurements

Understanding measurement uncertainty and machine variability has a significant role in any performance analysis, particularly when comparing models running across different platforms without advanced performance tools or methods like tracing or sampling. Before starting the collection

Table 1. List of CPMIP metrics collected.

Metric	Used to evaluate
Resolution (Resol)	Number of grid points $NX \times NY \times NZ$ per component
Complexity (Cmplx)	Number of prognostic variables per component
Platform	Machine measurements: core count, clock frequency, and double-precision operations per clock cycle
Simulation years per day (SYPD)	Number of simulated years per day (24 h) of execution time
Core hours per simulated year (CHSY)	Cost, measured in core hours per simulated year
Actual SYPD (ASYPD)	How queue time and interruptions affect the complete experiment duration
Parallelisation (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

of the metrics, we asked each institution to indicate the machine variability, which was reported to be below 10 % for all machines used. This provides an initial rough estimation, subject to future refinement efforts like the usage of benchmarking codes for climate science like the one proposed by van Werkhoven et al. (2023).

It is important to note that not all metrics exhibit the same variability range. Certain metrics, such as parallelisation, resolution, platform, and model complexity, are constant values determined just by the experimental configurations, the HPC infrastructure, and model characteristics. These are considered *static* metrics.

The rest of the metrics are related to the execution speed and are therefore subject to different sources of variability. On the one hand metrics like the SYPD or CHSY are well known by the community and straightforward to collect: this results in less margin of error during collection, and any variability should be attributed solely to the machine. On the other hand metrics like the actual SYPD, JPSY, coupling cost, memory bloat, data intensity, and data output cost are less common to collect, and this can lead to confusion and human errors (e.g. whether the actual SYPD should include system interruptions or only queue time can lead to systematic misreporting). This represents a second source of variability, difficult to assess and estimate.

Identifying and understanding this uncertainty is key for accurately interpreting and comparing the performance of models across different centres. Special effort has been made to ensure the quality and correctness of the metrics presented in this work by providing continuous support to the groups during data collection and double-checking the reported numbers with the responsible parties of each institution whenever necessary.

Future collections like this one will contribute to better identification and addressing of metric uncertainty, while detailed analysis of individual metrics will enhance our understanding of their characteristics and exhibit variability. For instance, studies like Acosta et al. (2023) focus mainly on the coupling cost and offer valuable lessons for understanding and measuring this metric, therefore mitigating possible uncertainties arising from misconception or lack of appropriate tools to collect them in the future.

3 Analysis

Analysing 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 Sect. 2: in Tables 2 and 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 of the run, the resolution used for the OCN and ATM, the reference for the experiment configuration, and the CPMIP metrics. Additionally, we include in Table 4 the most relevant information on the HPC platforms used by the institutions and some supplementary metrics in Table 9 related to the execution costs in CO_2 emissions.

Later, for each of the metrics analysed in detail in the following sections, we filtered by model, selecting those where the metric was provided and 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.

Table 2. List of institutions and models that provided the metrics from their CMIP6 executions. Also listed are the HPC platform and resolution used for the atmosphere (ATM) and ocean (OCN) components. Note that "resol" in Table 1 is defined as the number of grid points. For better readability, we present here this information using the more conventional measurement of degrees of latitude and longitude.

Institution	Experiment	HPC machine	Atmosphere resol	Ocean resol	Reference
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	0.25	
	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	0.25	
	OM4-p25			0.25	
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	0.25	
	HadGEM3-GC3.1-HH		0.8	0.08	
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	0.25	

3.1 Resolution

We initially attempt to extract valuable information from Table 3 by grouping the experiments based on resolution. This allows for a comparison of the performance achieved by ESMs with similar targets. 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 ($NX \times NY \times NZ$), and the total resolution is given by the sum of the resolutions of their constituents. There is no strict consensus on the connection between the number of grid points and the categorisation 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 configu-

https://doi.org/10.5194/gmd-17-3081-2024

Table 3. List of institutions with the model and CPMIP metrics. We also include the useful simulated years (useful SYs), which account for the number of years simulated by each experiment that generated data with scientific value.

Institution	Experiment	Resol	Cmplx	SYPD	ASYPD	CHSY	Paral	JPSY	Cpl C	Mem B	DO	DI	Useful SYs
BSC	EC-Earth3	1.99×10^7	34	15.20	9.87	1213	768	4.41×10^7	0.080	59.5	1.12	0.041	14 020
	EC-EarthVeg	1.99×10^7		12.36	7.42	1491	768	4.87×10^7	0.100	68.48	1.13	0.059	252
CMCC	CM2-SR5	6.94×10^6	397	6.68	6.50	2069	576	1.67×10^9	0.074	17.8	1.04	0.050	965
CNRM-CERFACS	CNRM-CM6-1-atm	2.98×10^6	128	7.30	6.10	1292	393	3.50×10^7					5723
	CNRM-CM6-1	1.02×10^7	181	8.10	7.30	1352	400	3.38×10^7					22 241
	CNRM-CM6-1-HR-atm	2.36×10^7	128	2.20	1.80	1541	520	4.80×10^7					1190
	CNRM-CM6-1-HR	1.37×10^8	181	1.50	1.48	4289	840	1.07×10^8					1642
	ESM2-1-atm	2.98×10^6	335	7.10	6.40	8520	781	2.28×10^8					1759
	ESM2-1	1.10×10^7	393	4.70	4.40	21 552	1347	5.28×10^8					11 761
DKRZ	MPI-ESM1-HR	2.00×10^7		13.33	11.00	4710	2616	3.21×10^8					1864
GFDL	OM4-p5	3.32×10^7	13	15.90	12.22	1962	1300	7.50×10^7	0.140	33.61		0.039	300
	ESM4-piC	3.76×10^7	140	8.65	7.46	13 576	4893	5.19×10^8	0.270	40.57		0.003	1124
	CM4-piC	1.28×10^8	31	9.98	8.16	15 388	6399	3.72×10^8	0.130	47.64		0.018	657
	OM4-p25	1.26×10^8	11	11.50	7.05	9748	4671	5.88×10^8	0.260	16.09		0.006	300
IITM	IESM	1.83×10^6	168	8.00	7.00	996	332	3.81×10^7		36.7			845
IMPE	BESM	6.88×10^6	132	3.60	3.40	1853	278					0.020	360
IPSL	IPSL-CM6A	1.06×10^7	750	12.00	11.50	1900	950	1.16×10^8	0.050	10.00	1.20	0.070	53 000
KNMI	EC-Earth3	1.99×10^7	34	16.20	16.20	1286	868						1009
	EC-Earth3-AerChem	2.06×10^7		3.03	3.03	3549	448						730
MPI	MPI-ESM1-LR-ATM	8.66×10^5		45.90	25.20	163	312	1.11×10^7					991
	MPI-ESM1-LR-LAND	8.33×10^5		282.80	265.40	3	36	1.39×10^5					2460
	MPI-ESM1-LR	3.12×10^6		55.60	22.70	379	878	2.56×10^7					18 860
NERC	UKESM1-AMIP	2.35×10^6	202	1.64	1.41	7376	504	1.04×10^8		52.50	1.31	0.003	45
	UKESM1-0-LL	1.14×10^7	252	2.02	1.10	8554	720	3.18×10^8	0.078	28.00	1.19	0.005	195
	HadGEM3-GC3.1-LL	1.14×10^7	150	4.25	1.06	12 198	2160	4.33×10^8	0.047	56.80	1.41	0.016	70
	HadGEM3-GC3.1-HM	1.99×10^8	54	0.58	0.46	192 662	4656	7.70×10^9	0.210	154.00		0.001	65
	HadGEM3-GC3.1-HH	1.26×10^9	54	0.49	0.34	588 931	12 024	2.30×10^{10}		183.00	1.41	0.0004	65
NorESM	NorESM2-LM	1.01×10^7		13.84	3.03	1665	960	5.60×10^7	0.035			0.065	5463
	NorESM2-MM	1.14×10^7		8.96	6.14	4886	1824	1.65×10^8	0.32			0.060	1021
SMHI	EC-EarthVeg	1.99×10^7		12.44	6.65	1667	864					0.028	6337
UKMO	HadGEM3-GC3.1-LL	1.14×10^7	228	4.00	3.55	13 392	2232	4.97×10^8	0.061	46.00	1.03	0.074	5610
	UKESM1-0-LL	1.14×10^7	372	4.30	3.60	16 074	2880	5.97×10^8	0.098	4.60	1.03	0.019	15 435
	HadGEM3-GC3.1-MM	1.44×10^8	236	1.65	1.32	62 836	4320	2.33×10^9	0.105	120.00	1.02	0.050	2386

Table 4. List of HPC machines used for the experiments under study, detailing hardware specifications, benchmark results (Linpack and high-performance conjugate gradient, HPCG), theoretical performance (Rpeak), power consumption, and power usage effectiveness (PUE) for each system. PFlops: petaflops; TFlops: teraflops.

Institution	Machine	Total cores	Cores per node	Memory node (GB)	Memory per core (GB)	Network	CPU family	CPU frequency (GHz)	Rpeak (PFlops s^{-1})	Linpack (PFlops s^{-1})	Power (kW)	HPCG (TFlops s^{-1})	PUE
BSC	MN4	155 520	48	96	2.00	Intel Omni-Path	Platinum Skylake	2.10	10.300	6.22	1632	122.24	1.35
CMCC	Zeus	12 528	36	96	2.67	InfiniBand	Gold Skylake	3.00	1.202				1.84
CNRM-CERFACS	Beaufix2	73 440	40	64	1.60	InfiniBand	E5 Broadwell	2.20	2.590	2.16	830	35.34	
DKRZ/MPI	Mistral	100 200	30	68	2.25	InfiniBand	E5 Haswell	2.29	3.960	3.01	1116	44.11	1.19
IITM	AADITYA	38 144	16	64	4.00	InfiniBand	E5 Haswell	2.60	0.790	0.72	790		
INPE	xc50	4080	40	192	4.80	Aries Interconnect	Gold Skylake	2.40	0.313				
IPSL	Curie	80 640	16	64	4.00	InfiniBand	E5 Sandy Bridge	2.70	1.670	1.36	2132	50.99	1.43
IPSL	Irene	79 488	48	192	4.00	InfiniBand	Platinum Skylake	2.70	6.640	4.07	917	52.68	
KNMI	Rhino	4752	28	128	4.57	InfiniBand	Nehalem	3.06	0.058				
NERC	Archer xc30	118 080	24	64	2.67	Aries Interconnect	E5 Ivy	2.70	2.550	1.64		80.79	1.10
NorESM	Fram	32 256	32	64	2.00	InfiniBand	E5 Broadwell	2.10	1.100	0.95			
SMHI	Beskow	65 920	32	64	2.00	Aries Interconnect	E5 Haswell	2.30	2.440	1.80	842		
SMHI	Tetralith	61 056	32	96	3.00	Intel Omni-Path	Gold Skylake	2.10	4.340	2.97		65.24	
UKMO	xc40	241 920	36	192	5.33	Aries Interconnect	E5 Broadwell	2.10	8.130	7.04			1.35

rations have been categorised as low resolution and use up to 2.10×10^7 grid points in total or no less than 0.7° latitude–longitude grid spacing for any of the components (see Fig. 1 and Table 2). On the other hand, only those experiments with an ocean/atmosphere resolution under 0.5° are treated as medium–high-resolution configurations (see Fig. 2).

We see the low-resolution experiments in Fig. 1. The number of grid points for each model's and institution's ATM (red) and OCN (blue) components has been listed in ascending order. Except for EC-Earth, we observe that all other models run the OCN at a higher resolution than the ATM. More precisely, the OCN resolution is between 3 and 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 $2/3$ is 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 used for the ATM. As one would expect, the total number of grid points of an experiment can be mainly explained by the ATM and OCN resolution used, but we show later how adding more components and/or features (in yellow in Fig. 1) can have a significant impact on the performance as well.

Figure 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, the GFDL CM4-piC experiment uses 55 times more grid points for the OCN component). The ATM resolutions range between 1 and 0.4°, 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 $1/12$°.

3.2 Complexity

The complexity of a coupled model, as defined in Table 1, accounts for the number of prognostic variables among all components. Here, "prognostic" refers to variables that the model directly predicts, such as temperature, atmospheric humidity, and salinity, in other words, variables that can be obtained directly as outcomes of the model. This metric is not well known by the community and has never been collected before, leading to confusion in some cases. Therefore, the values reported have to be seen as approximations. Only by continuously measuring these metrics in future collections will we improve our understanding of model complexity and its implications for model performance. The data in Table 5 reveal a wide variability in *complexity* (*Cmplx*) across the models, with most models reporting a value that ranges between 100 and 400. Notably, GFDL (OM4 and CM4) and EC-Earth have considerably lower Cmplx. The 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 ex-

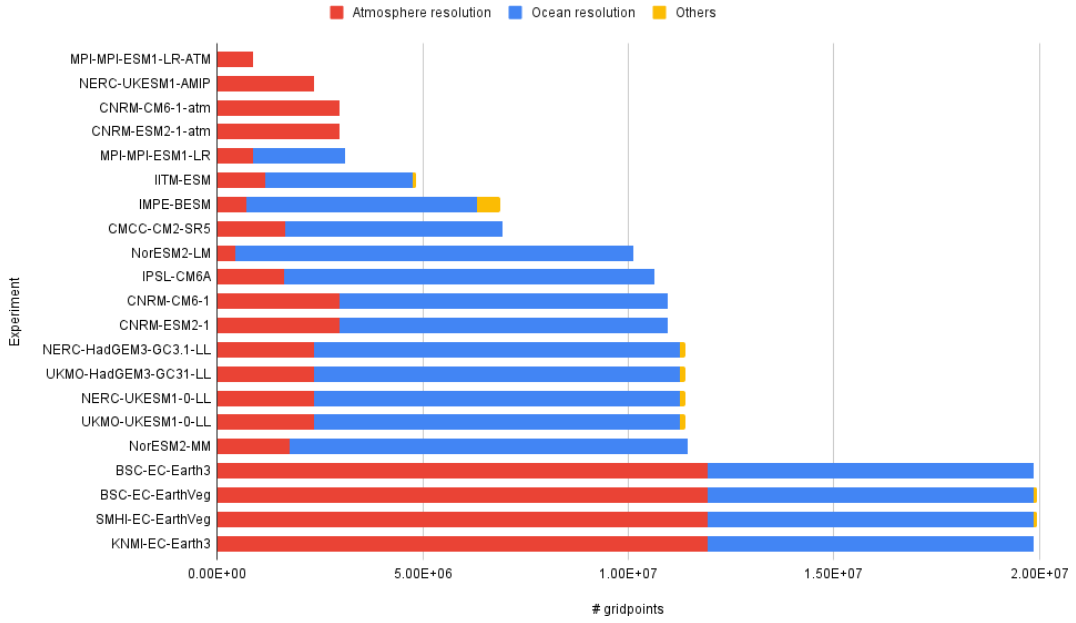

Figure 1. Atmosphere and ocean grid points for low-resolution experiments. The yellow colour refers to components that contribute to the atmosphere or the ocean but cannot be counted as a general circulation model per se (e.g. land surface, sea ice, vegetation).

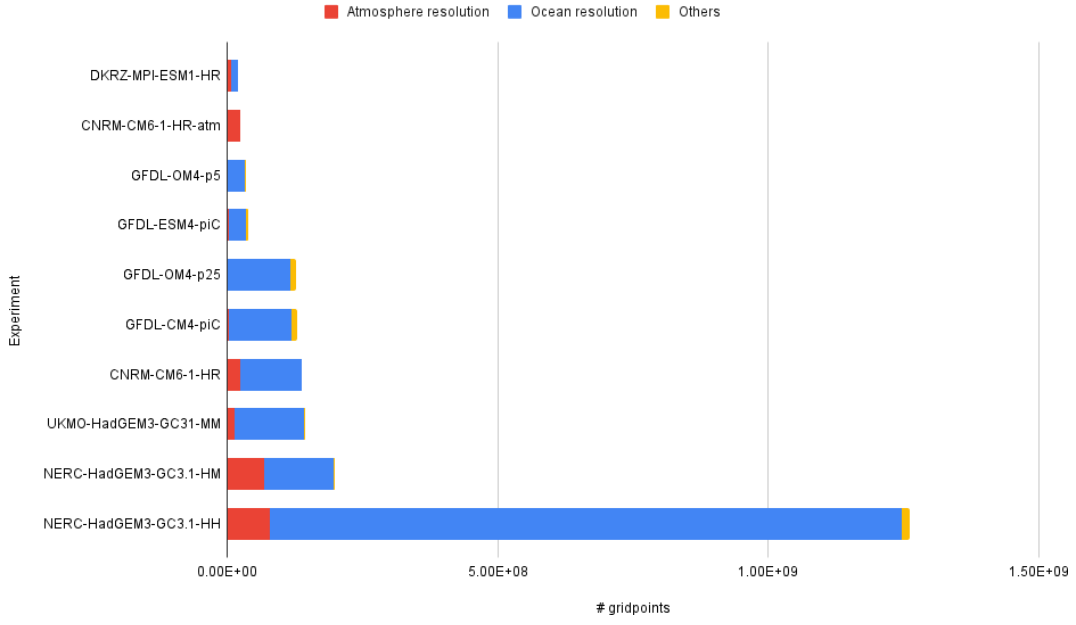

Figure 2. Atmosphere and ocean grid points for medium–high-resolution experiments.

pecting 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, the LPJ-Guess restart file size can measure tens of gigabytes and depends on the parallelisation used for this component. What is 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 dif-

Table 5. Resolution, SYPD, CHSY, Paral, and coupling cost for experiments that reported the complexity metric.

Institution	Experiment	Resolution	SYPD	CHSY	Parallelisation	Complexity	Coupling cost
BSC	EC-Earth3	1.99×10^7	15.20	1491	768	34	0.080
CNRM-CERFACS	CNRM-CM6-1-atm	2.98×10^6	7.30	1292	393	128	
	CNRM-CM6-1	1.10×10^7	8.10	1541	520	181	
	CNRM-CM6-1-HR-atm	2.36×10^7	2.20	8520	781	128	
	CNRM-CM6-1-HR	1.37×10^8	1.50	21 552	1347	181	
	ESM2-1-atm	2.98×10^6	7.10	1352	400	335	
	ESM2-1	1.10×10^7	4.70	4289	840	393	
GFDL	OM4-p25	1.26×10^8	11.50	9748	4671	11	0.130
	OM4-p5	3.32×10^7	15.90	1962	1300	13	0.140
	CM4	1.28×10^8	9.98	15 388	6399	31	0.260
	ESM4	3.76×10^7	8.65	13 576	4893	140	0.270
IITM	IESM	1.83×10^6	8.00	996	332	168	
IMPE	BESM	6.88×10^6	3.60	1853	278	132	
IPSL	IPSL-CM6A	1.06×10^7	12.00	1900	950	750	0.050
KNMI	EC-Earth3	1.99×10^7	16.20	1286	868	34	
NERC	HadGEM3-GC3.1-HM	1.99×10^8	0.58	192 662	4656	54	0.210
	HadGEM3-GC3.1-HH	1.26×10^9	0.49	588 931	12 024	54	
	HadGEM3-GC3.1-LL	1.14×10^7	4.25	12 198	2160	150	0.047
	UKESM1-AMIP	2.35×10^6	1.64	7376	504	202	
	UKESM1-0-LL	1.14×10^7	2.02	8554	720	252	0.078
UKMO	HadGEM3-GC31-LL	1.14×10^7	4.00	13 392	2232	228	0.061
	HadGEM3-GC31-MM	1.44×10^8	1.65	62 836	4320	236	0.105
	UKESM1-0-LL	1.14×10^7	4.30	16 074	2880	372	0.098

ferences between NERC and UMKO measurements, despite both running HadGEM-GC3.1 and UKESM1 models but on different platforms, raise questions about their source, which requires further investigation.

Nonetheless, the data from CNRM-CERFACS provide 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 processing elements when running the coupled version. This shows a clear interconnection between the parallelisation 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 the 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 synchronisations and interpolations (see GFDL results in Table 5). The relation between Cmplx and the coupling cost is further discussed in Sect. 3.5.

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 can 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. gigabytes per core hour).

3.3.1 Data output cost

From Table 6, we see that all the experiments conducted by UKMO and CMCC reported a data output cost below 1.05, even though the data intensity varies considerably between the different experiments. Moreover, we observe that the data output cost 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 the specific model's I/O configuration in influencing the data output cost metric. Besides, neither the metrics collected from UKMO nor the ones reported from NERC show that the data output cost should increase when running higher-resolution experiments (HadGEM3-GC31-MM and HH configurations). Moreover, EC-Earth and EC-Earth-Veg data output cost measurements are almost the same, suggesting that adding the vegetation model to EC-Earth does not increase the cost of the I/O, while UKESM runs conducted by NERC show that the data output cost is much higher when running the ATM standalone configuration, UKESM-AMIP, than the coupled run, UKESM-1-LL. Thus, the increase in complexity or resolution does not increase the cost of the I/O but the cost of the whole ESM simulation, which can diminish the data output cost metric if the I/O workload stays constant.

3.3.2 Data intensity

As seen in Table 6, the data intensity is generally of the order of megabytes 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 data intensity 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 data intensity.

3.4 Workflow and infrastructure costs

The real execution time of climate experiments cannot 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 and/or workflow manager, etc. can heavily extend the time to solution of ESMs. From the data in Table 1, we see that the difference between the reported SYPD and ASYPD varies a lot between institutions. Some claim that they had no overhead in their runs (KNMI), while for others, it can account for 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.

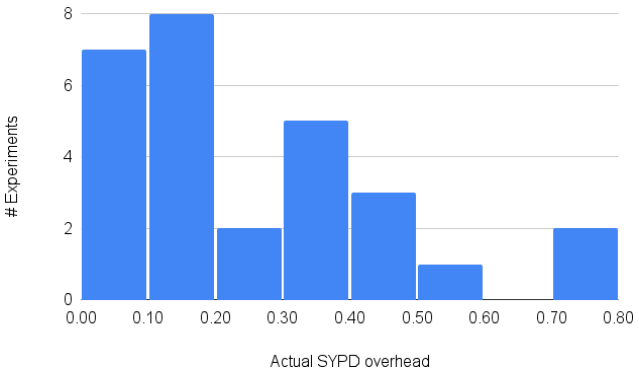

Figure 3. Histogram of the actual SYPD overhead.

The results support the idea that queuing time represents an increment of around 10 %–20 % of the speed of the ESM. On the other hand, if interruptions and workflow management overhead are also included, the total execution time can grow by up to 40 %–50 % compared to 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. From the metrics listed in Table 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 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 approximately 4 SYPD, but totally dif-

Table 6. Experiments that reported the data output cost (DO) and data intensity (DI) metrics.

Institution	Experiment	Resolution	Complexity	SYPD	CHSY	Parallelisation	DO	DI
BSC	EC-Earth3	1.99×10^7	34	15.20	1213	768	1.12	0.0410
	EC-EarthVeg	1.99×10^7		12.36	1491	768	1.13	0.0590
CMCC	CM2-SR5	6.94×10^6	397	6.68	2069	576	1.04	0.0500
GFDL	OM4-p5	3.32×10^7	13	15.90	1962	1300		0.0392
	OM4-p25	1.26×10^8	11	11.50	9748	4671		0.0178
	ESM4-piC	3.76×10^7	140	8.65	13 576	4893		0.0032
	CM4-piC	1.28×10^8	31	9.98	15 388	6399	1.24	0.0058
IMPE	IMPE-BESM	6.88×10^6	132	3.60	1853	278		0.0200
IPSL	IPSL-CM6A	1.06×10^7	750	12.00	1900	950	1.20	0.0700
NERC	HadGEM3-GC3.1-LL	1.14×10^7	150	4.25	12 198	2160	1.41	0.0160
	HadGEM3-GC3.1-HM	1.99×10^8	54	0.58	192 662	4656		0.0006
	HadGEM3-GC3.1-HH	1.26×10^9	54	0.49	588 931	12 024	1.41	0.0004
	UKESM1-AMIP	2.35×10^6	202	1.64	7376	504	1.31	0.0030
	UKESM1-0-LL	1.14×10^7	252	2.02	8554	720	1.19	0.0050
NorESM	NorESM2-LM	1.01×10^7		13.84	1665	960		0.0650
	NorESM2-MM	1.14×10^7		8.96	4886	1824		0.0600
SMHI	EC-EarthVeg	1.99×10^7		12.44	1667	864		0.0280
UKMO	UKESM1-0-LL	1.14×10^7	372	4.30	16 074	2880	1.03	0.0190
	HadGEM3-GC31-LL	1.14×10^7	228	4.00	13 392	2232	1.03	0.0740
	HadGEM3-GC31-MM	1.44×10^8	236	1.65	62 836	4320	1.02	0.0500

ferent 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 parallelisation, the smaller the overhead due to the workflow.

3.5 Coupling cost

Coupling cost (Eq. 3) 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 processing elements idle. It is defined as follows:

$$\text{Coupling\_Cost} \equiv \frac{T_\text{M} P_\text{M} - \sum_c T_\text{C} P_\text{C}}{T_\text{M} P_\text{M}}, \tag{3}$$

where T_M and P_M are the runtime and parallelisation for the whole coupled model, and T_C and P_C are the same for each individual component it uses.

Figure 4 shows the list of institutions ordered from lower to higher coupling cost. Most institutions reported that the cost increase due to the coupling accounts for around 5 %–15 % of the total. Only 4 (of 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) suggest that the increase in complexity leads to higher coupling cost 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 do not 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 coupling cost, as discussed in more detail in Sect. 3.2.

In general, the coupling cost tends to rise when running experiments that use a higher parallelisation. 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 coupling cost 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, etc.) not only for the coupled version but also per component.

3.6 Speed, cost, and parallelisation

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, *parallelisation* (i.e. the number of parallel resources allocated) stands out as a factor closely related to model speed and directly influences the computational cost (CHSY) of the model execution. This section shows 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 parallelisation allocated. However, if we analyse how the same model performs on different HPC machines (Table 7), we note that higher values of parallelisation usually correspond to faster but more energy-consuming simulations.

As seen in Fig. 5, the Paral and the CHSY are closely correlated in low-resolution models (e.g. CMCC-CM2-SR5, NorESM2-LM, IPSL-CM6A, NERC-HadGEM3-GC3.1-LL, UKMO-HadGEM3-GC3.1-LL, UKMO-UKESM1-0-LL, NorESM2-MM, BSC-EC-Earth3, BSC-EC-EarthVeg, KNMI-EC-Earth3, SMHI-EC-EarthVeg), showing that models do not scale in the current generation of HPC platforms. Otherwise, one would see that the CHSY of ESMs with similar resolution do not increase when using more processors given that the models run faster (i.e. higher SYPD). From the data, it is also clear which models are underperforming. Take for instance KNMI-EC-Earth3-AerChem, which, despite using a smaller parallelisation compared to its family counterparts (BSC-EC-Earth3, BSC-EC-EarthVeg, KNMI-EC-Earth3, and SMHI-EC-EarthVeg), exhibits a higher CHSY. Similarly, NERC-UKESM1-AMIP and NERC-UKESM1-0-LL employ less parallelisation compared to UKMO-UKESM1-0-LL, yet the CHSY does not decrease proportionally. Also, as illustrated in Fig. 6, the level of parallelisation 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. Future collections that include more medium–high-resolution experiments will help in creating further relationships for these experiments.

In addition, and as already discussed in Sect. 3.5, the coupling cost grows together with the parallelisation, although there is no sign that it limits the speed of the models.

3.7 Memory bloat

The memory bloat (Eq. 4) is the only CPMIP metric to evaluates models' memory usage by computing the ratio between the real and the ideal memory size. It is defined as

$$\text{Memory\_Bloat} \equiv \frac{M - \text{Parallelisation} \cdot X}{M_\text{i}}, \tag{4}$$

where M is the actual memory size, X is the binary file size, and M_i is the ideal memory size. The ideal memory size represents the size of the complete model state, which can be obtained by exploring the restart file size. This ratio typically falls between 10–100. Large memory bloat values signal excessive buffering. As an example (Balaji et al., 2017), for a rectangular grid with a halo size of 2 in the x and y directions and a 20×20 domain decomposition, the 2-D array including halos is 576 (24×24) instead of 400 (20×20), resulting in a bloat factor of 1.44. Similarly, a 10×10 decomposition would yield an array area of 196 and a bloat ratio of 1.96.

Table 8 presents the memory bloat values reported for various models along with other CPMIP metrics. We observe how the memory bloat increases with the resolution (e.g. NERC-HadGEM31), likely due to larger subdomains assigned to each compute unit in higher resolutions if the parallelisation does not increase proportionally. Additionally, memory bloat also increases when complexity grows, and the parallelisation remains constant (e.g. BSC-EC-Earth3 with and without the vegetation model) as it requires keeping more data in memory. It is important to acknowledge the challenges in obtaining accurate memory usage for such applications, and the authors are aware that institutions faced difficulties in providing these data. Therefore, the reliability of the reported values varies between sources and should be contrasted by future measurements (e.g. CPMIP collection for CMIP7). Precise memory measurements, however, can only be achieved with more advanced tools and approaches (memory profilers, MPI environment variables, etc.).

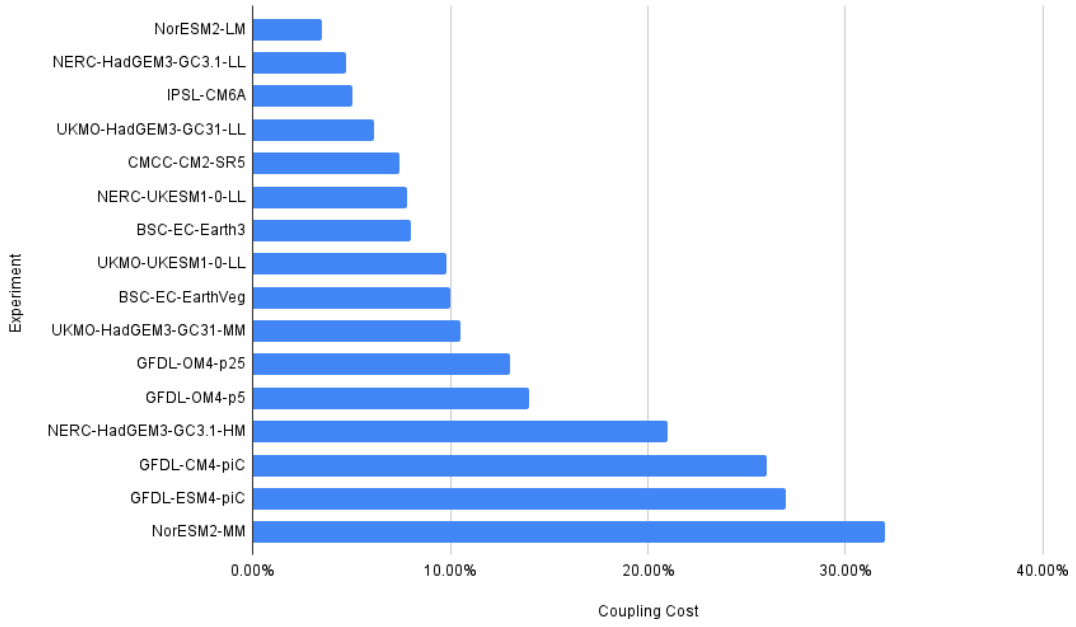

Figure 4. Coupling cost for all the institutions that provided the metric.

Table 7. Metrics for models available on different HPC platforms.

Experiment	Institution	Resolution	SYPD	CHSY	Parallelisation
EC-Earth3	BSC	1.99×10^7	15.20	1213	768
	KNMI	1.99×10^7	16.20	1286	868
EC-Earth3Veg	BSC	1.99×10^7	12.36	1491	768
	SMHI	1.99×10^7	12.44	1667	864
HadGEM3-GC3.1-LL	NERC	1.14×10^7	4.25	12 198	2160
	UKMO	1.14×10^7	4.00	13 392	2232
UKESM1-0-LL	NERC	1.14×10^7	2.02	8554	720
	UKMO	1.14×10^7	4.30	16 074	2880

3.8 Carbon footprint

In addition to the CPMIP collection, we have also gathered the general metrics shown in Table 9. 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 Table 9 to be a very good representation of the effort made for the collection during CMIP6. In any case, and taking into account the previous reasons, we do not analyse 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 kilowatt-hour-to-CO_2 conversion rates provided by each energy supplier, we calculated the carbon footprint (in tonnes of CO_2) using Eq. (2). As the reader can see, NERC reported 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 CO_2 emissions per kilowatt-hour. Regarding machine efficiency, CMCC reported that Zeus is the least power-efficient machine, with a

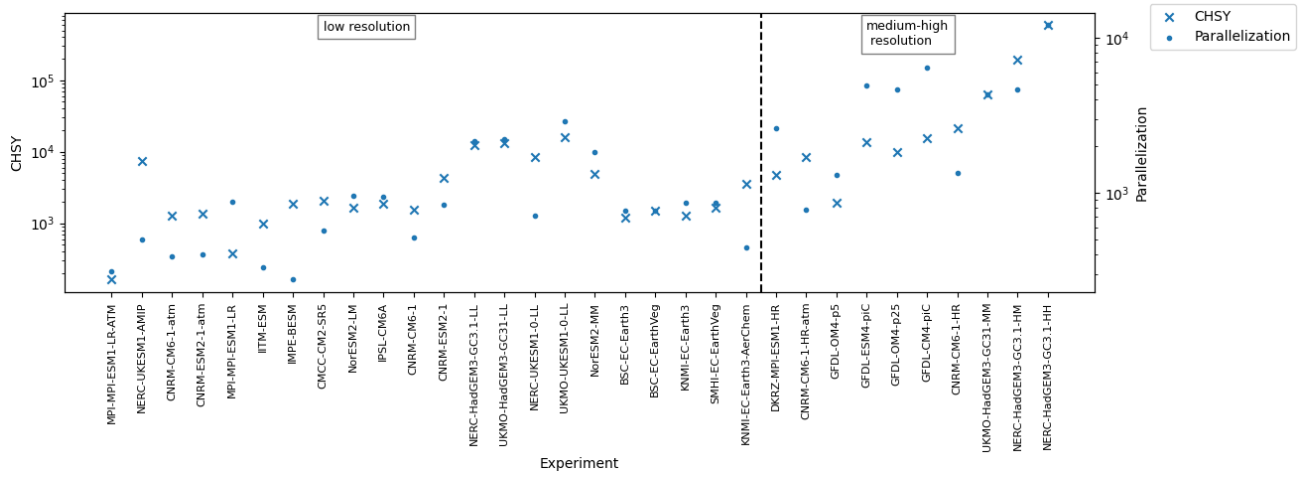

Figure 5. Comparison between CHSY and parallelisation for both low- and medium–high-resolution experiments. Experiment configurations are arranged from left to right in ascending number of grid points. Note that the vertical axis uses a logarithmic scale for better visualisation.

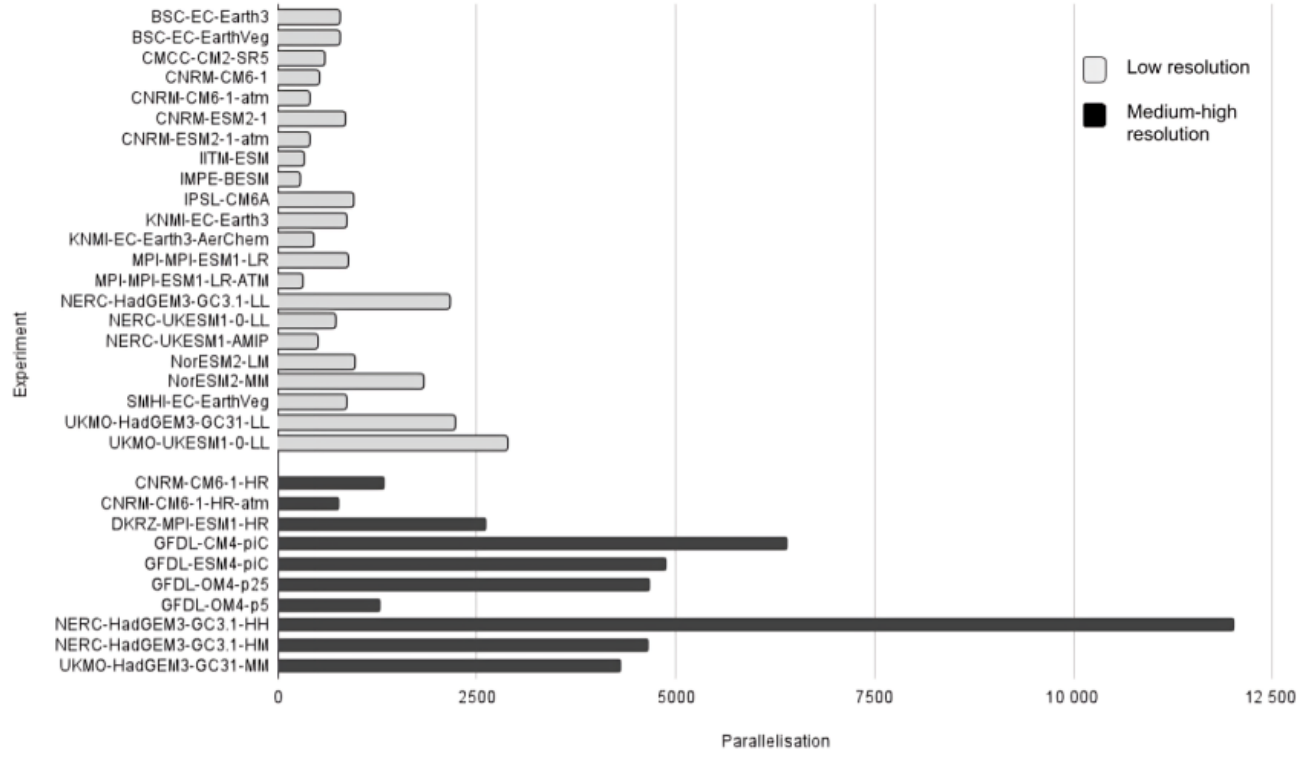

Figure 6. Parallelisation for low- (grey) and medium–high-resolution (black) models.

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 less than 1000 years yet reported nearly double the CO_2 emissions compared to EC-Earth or IPSL, despite these institutions having simulated longer experiments (in simulated years, SYs). The total energy cost of UKMO seems too high compared to the reported useful simulated years. 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 t CO_2, even when accounting for the experiments executed by only 8 out of the 49 institutions that are enlisted in CMIP6 (WCRP CMIP6 controlled vocabularies: https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_institution_id.html, last access: 2 February 2024). Based on a 2018 study by Acosta and Bretonnière (2018), the Earth Sci-

Table 8. Resolution, SYPD, CHSY, parallelisation, and memory bloat results for UKESM, EC-Earth, and HadGEM3-GC31 experiments.

Experiment	Resolution	SYPD	CHSY	Parallelisation	Memory bloat
BSC-EC-Earth3	1.99×10^7	15.2	1213	768	59.50
BSC-EC-EarthVeg	1.99×10^7	12.4	1491	768	68.48
NERC-HadGEM3-GC3.1-LL	1.14×10^7	4.3	12 198	2160	56.80
NERC-HadGEM3-GC3.1-HM	1.99×10^8	0.6	192 662	4656	154.00
NERC-HadGEM3-GC3.1-HH	1.26×10^9	0.5	588 931	12 024	183.00
UKMO-HadGEM3-GC31-LL	1.14×10^7	4.0	13 392	2232	46.00
UKMO-HadGEM3-GC31-MM	1.44×10^8	1.7	62 836	4320	120.00
NERC-UKESM-AMIP	2.35×10^6	1.6	7376	504	52.50
NERC-UKESM-LL	1.14×10^7	2.0	8554	720	28.00

Table 9. Other CMIP6 measurements. The "useful" metric, whenever used, accounts only for experiments that led to scientific value. The power usage effectiveness (PUE) depends on the HPC machine used (Table 4). The metric "person/months" quantifies the amount of work contributed by each institution to run the simulations, calculated as the product of the number of full-time personnel and the duration in months.

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 (TJ)	PUE	Conversion factor (MWh – kg CO_2eq)	Carbon footprint (t CO_2)
CMCC	965		0.097		1.99		7	1.61	1.84	408	329
CNRM-CERFACS	47 000		1.350	2.48	160.00	365.00	450	6.18	1.43	40	97
DKRZ	1276	1321	0.600		5.52	5.90		0.41	1.19	184	24
EC-Earth	28 105	38 854	0.800	1.41	31.13	46.36	115	1.24	1.35	357	165
IPSL	75 000	165 000	1.800	7.60	150.00	320.00	200	8.72	1.43	50	172
MPI-M	24 175	35 000	1.930		16.31			0.62	1.19	184	37
NCC-NorESM2	23 096		0.600		27.23	80.00	150	1.69			
NERC	640		0.460		55.50			2.17	1.10	0	0
UKMO	37 237		10.400		683.00			26.70	1.35	87	868

* The useful simulated years column values can differ from Table 1 given that some of the experiment runs are not shown in that table.

ence Group at the BSC, comprising around 80 people, had a CO_2 equivalent of commuting (29 t CO_2eq per year), computing infrastructure (397 t CO_2eq per year), building and infrastructure (117 t CO_2eq per year), and travel (255 t CO_2eq per year). The total budget was, therefore, estimated to be near 800 t CO_2eq per year. Consequently, the carbon footprint from the execution of only this small subset of experiments more than doubles our budget in a single year. This finding is consistent with observations from other groups within the community, such as a similar study conducted by CERFACS between 2019 and 2021, which reported a total budget of around 700 t CO_2eq per year. Nonetheless, the contributions that CMIP6 has made to climate science are invaluable and beyond the immediate costs associated with running the simulations.

4 Drawbacks and actions recommended

Thanks to the experience learned from the data collection and analysis done, we recognise 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 10 shows a list of the main drawbacks along with suggested actions for improvement.

5 Conclusions

One of the limiting factors for climate science is the computational performance that Earth system models (ESMs) can achieve on modern high-performance computing (HPC) platforms. This limitation imposes constraints on the number of years that can be simulated, the number of ensembles that can be used, the resolution used by the models, the number of features simulated in one experiment, I/O 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 visualise and understand the profiling informa-

Table 10. 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 are available. Integrating the CPMIPs in 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.	Normalise metric collection methods across institutions before multi-model runs. Develop tools to automatise 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.

tion, 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, and representative of the actual performance of ESMs and of the entire life cycle of modelling (i.e. simulation and workflow costs).

This paper presents, for the first time, the results obtained from the CPMIP collection during the CMIP6 exercise. It provides a list of the 14 institutions involved, primarily from the IS-ENES3 consortium, along with the 33 CMIP6 experiment configurations and the CPMIP metrics collected for each experiment. Furthermore, it goes well beyond mere data presentation and offers an in-depth analysis for each metric collected to demonstrate the broader utility of the CPMIP collection. For instance, this study investigates the resolution used by each model on the oceanic and atmospheric components; explores the relationship between execution speed and cost with the other metrics; assesses the impact of running models with higher processor counts, complexity, or I/O requirements; examines the overhead caused by queuing and workflow management; and explores the coupling cost across different configurations.

Besides the CPMIP metric analysis, this paper highlights results obtained from collaborations with other groups, such as the Carbon Footprint Group. This collaboration underscores the shared concern of multiple institutions regarding 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, including the heterogeneity of the models and HPC machines used, as well as uncertainty in the metric measurements reported. 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 the community can adopt for the development of novel tools and more finely grained metrics that will facilitate upcoming similar works. Moreover, the improvement and development of benchmarks specially designed for climate science will significantly enhance multi-platform performance comparisons. Continuous collection of these metrics in future multi-model projects (e.g. CMIP7) will facilitate the development of a shared database for the community and vendors.

Code and data availability. The original data used during this work can be found here: http://bit.ly/3Y6XhHM (Google Sheets, 2024). No software was used in this study.

Author contributions. MCA led the data collection process, ensuring that multiple institutions provided the necessary data and providing technical assistance throughout. SeP conducted the data analysis and validation and took the lead in writing the manuscript. SVPT 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. At least one of the (co-)authors is a member of the editorial board of *Geoscientific Model Development*. The peer-review process was guided by an independent editor, and the authors also have no other competing interests to declare.

Acknowledgements. The research leading to these results has received funding from the EU H2020 ISENES3 and co-funding from the Spanish National Research Council through OEMES.

Financial support. This research has been supported by the Barcelona Supercomputing Center (grant nos. PID2020-116324RA-I00 (OEMES) and H2020-GA-824084 (ISENES)).

Review statement. This paper was edited by Tatiana Egorova and reviewed by two anonymous referees.

References

Acosta, M. and Bretonnière, P.-A.: Towards Minimising Carbon Footprint of Climate Modelling: Modelling Centre Perspective, C report, 2018.

Acosta, M. C., Palomas, S., and Tourigny, E.: Balancing EC-Earth3 Improving the Performance of EC-Earth CMIP6 Configurations by Minimizing the Coupling Cost, Earth Space Sci., 10, e2023EA002912, https://doi.org/10.1029/2023EA002912, 2023.

Balaji, V.: Climate Computing: The State of Play, Comput. Sci. Eng., 17, 9–13, https://doi.org/10.1109/MCSE.2015.109, 2015.

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, Geosci. Model Dev., 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, Nat. Comput. Sci., 1, 104–113, https://doi.org/10.1038/s43588-021-00023-0, 2021.

Bondyopadhyay, P.: Moore's law governs the silicon revolution, Proc. IEEE, 86, 78–81, https://doi.org/10.1109/5.658761, 1998.

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, J. Adv. Model. Earth Sy., 12, e2019MS002010, 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, Int. J. High Perform. C., 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, Geosci. Model Dev., 15, 2973–3020, https://doi.org/10.5194/gmd-15-2973-2022, 2022.

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, J. Adv. Model. Earth Sy., 12, e2019MS002015, https://doi.org/10.1029/2019MS002015, 2020.

Frank, D., Dennard, R., Nowak, E., Solomon, P., Taur, Y., and Wong, H.-S. P.: Device scaling limits of Si MOSFETs and their application dependencies, Proc. IEEE, 89, 259–288, https://doi.org/10.1109/5.915374, 2001.

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.

Google Sheets: CPMIP metrics GMD, Google Sheets [data set], http://bit.ly/3Y6XhHM, last access: 23 February 2024.

Joussaume, S.: IS-ENES: Infrastructure for the European Network for Earth System Modelling, in: EGU General Assembly Confer-

ence Abstracts, EGU General Assembly Conference Abstracts, p. 6039, 2010.

Krishnan, R., Swapna, P., Vellore, R., Narayanasetti, S., Prajeesh, A. G., Choudhury, A. D., Singh, M., Sabin, T. P., and Sanjay, J.: The IITM earth system model (IITM ESM), in: Current Trends in the Representation of Physical Processes in Weather and Climate Models, Springer Singapore, 183–195, https://doi.org/10.1007/978-981-13-3396-5_9, 2021.

Liu, Z., Wang, B., Wang, T., Tian, Y., Xu, C., Wang, Y., Yu, W., Cruz, C. A., Zhou, S., Clune, T., and Klasky, S.: Profiling and Improving I/O Performance of a Large-Scale Climate Scientific Application, in: 2013 22nd International Conference on Computer Communication and Networks (ICCCN), 30 July–2 August, Nassau, Bahamas, 1–7, https://doi.org/10.1109/ICCCN.2013.6614174, 2013.

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), J. Adv. Model. Earth Sy., 14, e2021MS002814, https://doi.org/10.1029/2021MS002814, 2022.

McGuffie, K. and Henderson-Sellers, A.: Forty years of numerical climate modelling, Int. J. Climato., 21, 1067–1109, 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), J. Adv. Model. Earth Sy., 10, 1383–1413, https://doi.org/10.1029/2017MS001217, 2018.

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, J. Adv. Model. Earth Sy., 11, 4182–4227, https://doi.org/10.1029/2019MS001791, 2019.

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, Geosci. Model Dev., 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, J. Adv. Model. Earth Sy., 12, e2019MS001946, https://doi.org/10.1029/2019MS001946, 2020.

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, Zenodo [code], https://doi.org/10.5281/zenodo.7671032, 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, Geosci. Model Dev., 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, J. Adv. Model. Earth Sy., 11, 2177–2213, https://doi.org/10.1029/2019MS001683, 2019.

Wang, D. and Yuan, F.: High-Performance Computing for Earth System Modeling, 175–184, Springer International Publishing, Cham, ISBN 978-3-030-47998-5, 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, Comput. Sci. Eng., 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, J. Adv. Model. Earth Sy., 10, 357–380, https://doi.org/10.1002/2017MS001115, 2018.