IS-ENES3 Deliverable D3.1

Initial requirements on model evaluation

Reporting period: 01/07/2020 – 31/12/2021

Authors: Jérôme Servonnat (CNRS-IPSL), Kim Serradell (BSC), Klaus Zimmermann (SMHI), Rémi Kazeroni (DLR), Eric Guilyardi (CNRS-IPSL)
Reviewer(s): Valeriu Predoi (NCAS), Bouwe Andela (NLeSC), Christopher Kadow (DKRZ)
Release date: 30/06/2021

ABSTRACT

This deliverable reports on the results of the "*Community survey to review the model evaluation needs and expectations of a variety of end users both existing and future"* (Task 4.2 of NA/WP3) led by Assimila Limited. The main objectives of the study were to survey a wide range of stakeholders interested in Earth System Model (ESM) evaluation and hence to provide an evidence base for IS-ENES3 regarding the development of community ESM evaluation methods and tools.

Revision table			
Version	**Date**	**Name**	**Comments**
Release for internal review	03/06/2021	V1 draft	Content of the draft is directly copied from the Assimila report
Release for external review	09/06/2021	V1.1 draft	Contains the feedbacks of Kim Serradell
Final version after implementing comments from external reviewers	29/06/2021	V1 final	Contains the feedbacks from the reviewers: - Bouwe Andela - Valeriu Predoi - Christopher Kadow

Dissemination Level		
PU	Public	X
CO	Confidential, only for the partners of the IS-ENES3 project	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Executive Summary

This report has been prepared by Assimila Limited as an input to *D3.1 Initial requirements on model evaluation of Grant Agreement number: 824084 — Infrastructure for the European Network for Earth System modelling - Phase 3 (IS-ENES3)* and forms part of Task 4.2 of NA2/WP3: *Perform "Community survey to review the model evaluation needs and expectations of a variety of end users both existing and future"*. The work carried out is based on a scoping document from the IS-ENES3 team [RD1] and the Assimila proposal for the study [RD2]. The main objectives of the study were to survey a wide range of stakeholders interested in Earth System Model (ESM) evaluation and hence to provide an evidence base for IS-ENES3 regarding the development of community ESM evaluation methods and tools.

The work was carried out between May 2020 and May 2021, by surveying a broad section of the European user community, liaising closely with the IS-ENES3 Study Steering Group (see Appendix A). The survey method used open ended questions, mainly through one-to-one interviews. The interview guide was agreed with the Study Steering Group who also recommended interviewees. Transcripts of the interviews were agreed with participants and made available to the Steering Group to check for balance and completeness.

Key benefits of community tools identified by interviewees are:

- Reduced duplication of effort on often repeated tasks.
- Promotion of standardisation and hence enabling meaningful cross comparisons, eg between ESMs, ESMs and data. The World Climate Research Programme's Coupled Model Intercomparison Project (CMIP) has provided important impetus for progress in Earth System Models and their intercomparison by promoting standardisation: experiments, formats, naming conventions, diagnostics and metrics. In turn this has promoted development and use of evaluation tools.
- Critical mass to create a support and collaboration community in institutions and between institutions; the critical mass also helps to promote the sustainability of the tools, long term.
- More efficient use of available resources (funding, staff).

However, a number of interviewees also recognised the reality of custom in-house developments, based on:

- Freedom: Scientists like to do things their own way and need to be convinced to use "off-the-shelf" tools.
- Heritage: Force of habit and previous investment prevent convergence on common tools.

The points above set much of the context for the more detailed user requirements. There was good support, amongst project participants, for the continued development of a European community ESM evaluation tool. Whilst a diversity of approaches continues to stimulate innovation, ESMValTool has emerged over recent years as a well-regarded tool with many of the attributes and features required by a wide climate science community to carry out evaluations for model development and intercomparison and to support process studies. It is promoted by CMIP6 as a tool for model intercomparisons, including contributions to IPCC assessments.

There is sufficient momentum behind ESMValTool to recommend its continued development as a European community tool for ESM evaluation. Close liaison between data scientists, software developers and climate scientists will be essential to ensure a successful user driven approach. A long-term governance and funding strategy is needed to confirm its long term future with the climate user communities.

REFERENCE DOCUMENTS

<table>
<tr><th>Ref</th><th>Document</th><th>Description</th></tr>
<tr><td>RD1</td><td>IS-ENES3 NA2 Model evaluation survey
Draft 0.2, 15/05/19 following NA2 group review
Eric Guilyardi (IPSL), Kim Serradell (BSC), Axel Lauer (DLR), Klaus Zimmermann (SMHI)</td><td>IS-ENES3 NA2 Model evaluation survey
Draft 0.2, 15/05/19 following NA2 group review Eric Guilyardi (IPSL), Kim Serradell (BSC), Axel Lauer (DLR), Klaus Zimmermann (SMHI)</td></tr>
<tr><td>RD2</td><td>Community Survey of Model Evaluation Needs
Assimila Limited proposal B2019-02, Version 2.0, Dr Zofia Stott, 13 May 2020</td><td>Basis for the work carried out</td></tr>
<tr><td rowspan="2">RD3</td><td>Pan-European Survey of the Climate Modelling Community
Phase 1 Analysis and Synthesis Report
Assimila Project-2014-02, Version 2.0, 03/02/16</td><td>Final report on a study carried out in the context of the Joint Programming Initiative Climate to examine the potential to pool national research efforts in Earth System Modelling in order to make better use of Europe’s public R&D resources.</td></tr>
<tr><td colspan="2">http://www.jpi-climate.eu/media/default.aspx/emma/org/10875543/Survey_Final+report+on+European+ESM+V2.0.pdf</td></tr>
<tr><td></td><td></td><td></td></tr>
</table>

ACRONYMS AND ABBREVIATIONS

AEROCOM	International initiative on global aerosols
AI	Artificial Intelligence
ana4mips	Reanalysis for Model Intercomparisons Project
API	Application Programming Interface
BSC	Barcelona Supercomputing Center
CCI	Climate Change Initiative
CCMI	Chemistry Climate Model Initiative
CCMVal	Chemistry-Climate Model Validation Activity
CDO	Climate Data Operators
CEDA	Centre for Environmental Data Analysis
CF	Climate and Forecast
CLIMAF	Climate Model Assessment Framework
CMIP	Coupled Model Intercomparison Project
CMOR	Climate Model Output Rewriter
CNRM	Centre National de Recherches Météorologiques
CORDEX	Coordinated Regional Climate Downscaling Experiment
DECK	Diagnostic, Evaluation and Characterization of Klima
DKRZ	Deutsches Klimarechenzentrum (German Climate Computing Centre)
DLR	Deutsches Zentrum für Luft- und Raumfahrt (German Aerospace Center)
EMAC	ECHAM/MESSy Atmospheric Chemistry
ENSO	El Niño–Southern Oscillation
ESGF	Earth System Grid Federation
ESA	European Space Agency
ESM	Earth System Model
ESMValTool	Earth System Model Evaluation Tool
EU	European Union

FREVA	Free Evaluation System framework, Germany
GDFL	Geophysical Fluid Dynamics Laboratory
GDPR	General Data Protection Regulation
ghg	greenhouse gas
IPCC	Intergovernmental Panel on Climate Change
IPSL	Institut Pierre-Simon Laplace
IRIS	Met Office package for analysing and visualising Earth science data
IS-ENES3	Infrastructure for the European Network for Earth System Modelling3
MDTF	Model Diagnostics Task Force
MetO	Met Office
MJO	Madden–Julian Oscillation
NAO	North Atlantic Oscillation
NCL	NCAR Command Language
NCAS	National Centre for Atmospheric Science
NCO	netCDF Operator
NEMO	Nucleus for European Modelling of the Ocean
NetCDF	Network Common Data Form
NOAA	National Oceanic and Atmospheric Administration
obs4mips	Observations for Model Intercomparisons Project
PCMDI	Program for Climate Model Diagnostics and Intercomparison
PMP	PCMDI Metrics Package
R&D	Research and Development
SMHI	Sveriges Meteorologiska och Hydrologiska Institut (Swedish Meteorological and Hydrological Institute)
STFC	Science and Technology Facilities Council
UREAD	University of Reading

Table of contents

1. Objectives of the survey

1.1 Introduction

This report provides input to *D3.1 Initial requirements on model evaluation of Grant Agreement number: 824084 — Infrastructure for the European Network for Earth System modelling - Phase 3 (IS-ENES3)* and forms part of Task *4.2 of NA2/WP3: Perform "Community survey to review the model evaluation needs and expectations of a variety of end users both existing and future"*. The work carried out is based on a scoping document from the IS-ENES3 team [RD1] and the Assimila proposal for the study [RD2]. The main objectives of the study were to survey a wide range of stakeholders interested in Earth System Model (ESM) evaluation and hence to provide an evidence base for IS-ENES3 regarding community ESM evaluation methods and tools.

1.2 Background

[RD3], produced in the context of the Joint Programming Initiative Climate, looked at the activities of the main climate modelling groups in Europe and tried to identify areas where it made scientific and financial sense to pool national research efforts in order to overcome fragmentation, make better use of Europe's public R&D resources and hence, by facilitating cross border collaboration between top scientists, to tackle common European challenges in climate modelling more effectively.

One area which emerged as both consuming significant resources in all modelling centres and where increased collaboration could be effective was model evaluation. At the time of the study, progress had been made through the development of tools such as the Earth System Model Evaluation Tool (ESMValTool[1]) but model development stakeholders reported that more needed to be done. It was concluded that any programme of work in this area would need to be compatible with Coupled Model Intercomparison Project (CMIP) processes, and may require closer coordination with programmes such as the Copernicus Climate Change Service. Over time, it was felt that a coordinated model evaluation programme may also help to identify the most successful modelling approaches and so help to coordinate and target national resources in model development.

1.3 Scope of work carried out

Under the auspices of IS-ENES3, this study returned to the issue of model evaluation and looked in more depth at what is required, not only by model developers but also other stakeholders such as model users, data and infrastructure providers. It also gained insight from software developers

[1] https://www.esmvaltool.org/

of the evaluation tools themselves. The output of the study provides an understanding of the challenges faced by those developing model evaluation infrastructures and hence addresses key technical, standards, and governance requirements from the perspective of users and potential users.

At the outset of the study, there was an assumption that improved coordination of ESM evaluation was worthwhile in terms of accelerating scientific understanding and societal benefit from ESM activities, although no assumptions were made about the most effective way to implement a coordinated model evaluation programme. On the other hand, the study did not start from scratch – model evaluation tools are widely used. Given the development of ESMValTool as a European community tool over recent years, the study used ESMValTool as a "benchmark" to compare and contrast different approaches. This included understanding how ESMValTool is used today and gauging support and opportunities for the further development with a view to long-term adoption by a broad community at EU level. Equally, the views of users who have taken or promoted other approaches were equally valued.

2. Methodology

2.1 Introduction

The work was carried out between May 2020 and May 2021, by surveying a broad section of the European user community and liaising closely with the IS-ENES3 Study Steering Group (see Appendix A). The survey work was carried out mainly through structured interviews (using Microsoft Teams and similar online video conferencing technologies; initially planned to be held on site but eventually happened in visio conference because of restrictions due to Covid-19) using an interview guide agreed with the Steering Group and sent to participants in advance (see Appendix B). In total 41 requests for interviews were issued by the Steering Group and Assimila. 20 interviews were held, plus 5 additional email exchanges (see Appendix A for the list and the following section for the criteria to build the list). In addition:

- Information about the survey was disseminated to the ESMValTool User Engagement Team;
- Results were presented at EGU 2021, "*Model evaluation expectations of European ESM communities: first results from a survey",* Jérôme Servonnat, Eric Guilyardi, Zofia Stott, Kim Serradell, Axel Lauer, Klaus Zimmermann, Fanny Adloff, Marie-Claire Greening, Remi Kazeroni and Javier Vegas, on behalf of IS-ENES3.

- Results were also presented at a "*Virtual workshop on requirements for a fast and scalable evaluation workflow"* organised in the context of IS-ENES3 in close collaboration with the ESMValTool development team, 18-19 May 2021[2].

Edited transcripts of all interviews were iterated with, and agreed, with participants and made available to the Study Steering Group to gauge completeness and balance of the information collected.

The Study Steering Group met on 8 July 2020, 23 November 2020 and 9 March 2021 to review progress. A number of meetings were also held to prepare the presentation for EGU 2021.

2.2 Survey method

As shown in Appendix B, the survey method used open ended questions, mainly through one-to-one interviews. This method is generally employed for **qualitative** research and is very useful in understanding in detail the respondent and their position concerning a defined topic/situation. It is particularly helpful in revealing new aspects and issues that are unknown or unidentified at the outset.

The other common survey method is to provide closed-ended questions with a limited number of answers (yes/no, rank 1 to 5 etc) to a set of well-defined questions. This method is for quantitative research and allows standardised analysis of the data and facilitates prioritisation. It is useful to confirm the hypotheses (e.g. formulated following a survey with open-ended questions).

During this study the open-ended approach was used. It was agreed that in order to develop common ESM evaluation tools for the community, it was good practice to give potential users a say regarding their needs, without a set of preconceived ideas (closed questions). With a broad range of stakeholders, the method allowed interviewees to focus on aspects of most interest to them and explore ESM evaluation from their standpoint and experience. The interview guide provided a check list for each interview.

It may be useful to follow-up with a survey using closed-ended questions, for example if resource constraints on development tools necessitate prioritisation.

2 https://is.enes.org/events/workshops/is-enes3-virtual-workshop-on-requirements-for-a-fast-and-scalable-evaluation-workflow

2.3 Open-ended survey questions

The detailed survey questions, as sent to interview candidates are presented in Appendix B. In summary the questions aimed to understand:

- Context: job description of interviewee – to help focus on the questions likely to be of most interest to the interviewee.
- Which model evaluation tools the interviewee uses, is familiar with, or contributed to (including as developer, or data provider).
- How and why the tools are used, including issues of workflows and pros and cons of different approaches.
- Most important aspects of these evaluation tools and features used by the interviewee.
- The challenges of developing and using model evaluation tools.
- Future developments for the tools / challenges.
- Governance of a community tool with particular reference to ESMValTool.
- Other points arising.

2.4 Interviewees

The full list of study participants is provided in Appendix A. This list has been composed by the steering committee (authors of the study) together with Assimila at the beginning of the project. These were scientists and analysts interested in the following categories of using evaluation tools:

1. Model intercomparison diagnostics and metrics, particularly in the context of international developments such as CMIP, CORDEX and for IPCC.
2. Model development and benchmarking, particularly comparing successive model versions and comparing a model with data. The main motivation is to confirm model improvements with successive versions and to publish detailed analyses showing model improvements.
3. Understanding climate processes and demonstrating improved understanding through publication of diagnostics and metrics. Improved understanding, e.g. of El Nino, may then feed into subsequent model developments.
4. Preparing observational datasets for use with model outputs.
5. Providing information for climate change impact studies.
6. Development of tools.
7. Data science providing efficient access to model outputs and observation data.

Hence, in the context of this study, we interpret “ESM evaluation” broadly to encompass any or all of the activities above.

The study participants included:

- Senior scientists/professors, postdocs and industry representatives;
- Working scientists with direct hands-on interests in developing and using ESM evaluation tools and those with more strategic interests;
- Representatives from key institutions in France, Germany, Spain, Sweden, UK, USA

All interview responses were valued equally. Of the 28 participants in the study 13 had direct experience of either developing and/or using ESMValTool.

3. Results for community tools for ESM evaluation

3.1 Benefits of community tools for ESM evaluation

At the most fundamental level one issue raised by interviewees was whether community tool(s) are worthwhile. Key benefits identified by interviewees are:

- Reduced duplication of effort on often repeated tasks.
- Promotion of standardisation and hence enabling meaningful cross comparisons, e.g. between ESMs, ESMs and data. CMIP has provided important impetus for progress in Earth System Models and their intercomparison by promoting standardisation: experiments, formats, naming conventions, diagnostics and metrics. In turn this has promoted development and use of evaluation tools.
- Critical mass to create a sustainable support and collaboration community in institutions and between institutions.
- More efficient use of resources (funding, staff).

However, a number of interviewees also recognised the reality of custom in-house developments, based on:

- Freedom: Scientists like to do things their own way and need to be convinced to use “off-the-shelf” tools.
- Heritage: Force of habit and previous investment prevent convergence on common tools.

The points above set much of the context for the more detailed user requirements discussed below.

3.2 Overview of tools identified by interviewees

The study showed a range of “ESM evaluation tools” in use, including sophisticated diagnostic packages and frameworks such as:

- Auto-assess, UK Met Office in-house package (UK Met Office transitioning to ESMValTool)
- ESMValTool, community tool, coordinated and hosted by DLR, Germany
- PMP (PCMDI (Program for Climate Model Diagnosis and Intercomparison) Metrics Package), Lawrence Livermore National Laboratory, USA
- CLIMAF (Climate Model Assessment Framework), developed in France in context of model development and evaluation, IPSL, CNRM, CERFACS, France
- FREVA, Free Evaluation System Framework, Germany
- Model Diagnostics Task Force Diagnostics Package, Geophysical Fluid Dynamics Laboratory, NOAA, USA.

While a detailed comparison of these tools was not carried out during the study it was clear that they are not identical in terms of scope, use and governance, for example:

- Integrated into model development (eg Auto-assess at Met Office, CLIMAF at IPSL, CNRM)
- Used for "standard" multi-model analysis, eg CMIP6 DECK (ESMValTool, PMP) and IPCC assessments
- Supporting process understanding (specialised model outputs e.g. for precipitation or cloud processes to examine higher frequency phenomena, specialised metrics packages e.g. ENSO, MDTF Diagnostics Package).

There are also very useful tools for doing ESM analysis that are somewhat smaller in scope, such as:

- CDO (Climate Data Operators, Max Planck Institute, Germany) – a collection of command line Operators to manipulate and analyse Climate and NWP model Data. It can do basic operations for climate variables to, for example, calculate anomalies and more complex operations like re-gridding to compare different models or model-observations. From its command line base, scientists develop more elaborate shell scripts around this.
- NCO (NetCDF Climate operators) is another tool that allows operations on individual files and allows standardisation of climate data.
- NCL (NCAR Command Language) designed for scientific data analysis and visualisation.

These can be considered first generation tools. 2nd generation tools e.g. IRIS – a Python tool from the Met Office have a deeper understanding of the metadata which can benefit the user when considering the complexity of climate data. xarray and also the R-programming language are also included in these 2nd generation tools.

ESMValTool and others listed above as sophisticated diagnostic packages and frameworks can be considered 3rd generation tools, adding more complexity to 2nd generation tools. For example, ESMValTool has additional functionality in data handling - it can locate different types of global and regional climate data.

The study also took note of "toolboxes" such as the software environments being developed by the ESA Climate Change Initiative[3] and the Copernicus Climate Change Service[4] for ingesting, operating on and visualising climate data. Toolboxes cater for different needs and sophistication of users by integrating both command line and graphical user interfaces as well as Python API.

[3] http://climatetoolbox.io/

[4] https://cds.climate.copernicus.eu/toolbox/doc/index.html

3.3 Common user themes and needs

This section details what stakeholders participating in the study are looking for in community ESM evaluation tools.

GENREQ 1.0: Flexibility: tuned/tuneable to wide range of scientific needs

A community tool will need to cater for a wide range of users engaged in different aspects of climate science.

- Model development: Diagnostics are needed to compare different versions of a particular model and to compare simulations against observations as the model is developed. The suite of diagnostics and metrics needs to enable climate modellers to carry out standard checks at the early stages of model development together with more complex diagnostics and metrics as the model matures to provide confidence in the fidelity with which the model represent processes that affect climate and to understand model biases – described by one interviewee as needing to "get under the bonnet" of their latest climate model. The evaluation process is highly integrated with model development and production of climate simulations. The focus here is on the needs of one particular model. Since models are all different in many key scientific and technical aspects, modelling centres have tended to develop their own model evaluation tools to meet their own needs.
- Model analysis: Once a model has reached sufficient maturity, tailored, complex diagnostics and metrics are required for publications providing evidence of model improvement.
- Process studies: Process studies are designed to improve understanding of poorly described (or previously neglected) physical and biogeochemical processes that are central to the behaviour of the climate system. The ultimate goal is to improve the representation of the fundamental mechanisms in the Earth System Models. They often compare observational datasets with tailored numerical modelling efforts (e.g. focusing on processes such as ENSO) and thus intrinsically rely on evaluation tools.
- Model intercomparison: Enhanced scrutiny of models and expanded diagnostic analysis of model behaviour have been increasingly facilitated by internationally coordinated efforts to collect and disseminate output from model experiments performed under identical conditions. This has encouraged more comprehensive evaluation of models. The expanded evaluation effort, encompassing many different perspectives, makes it less likely that significant model errors are being overlooked and hence provides increased confidence in calculations of climate sensitivity and projections. With respect to evaluation, the technical emphasis here is on core pre-processing functions and standard diagnostics and metrics for a wide range of model outputs, observations and reanalyses to put them side to side (ultimately improving diagnosing and understanding the causes and effects of model biases and inter-model spread). The focus here is on outputs from many models, particularly as part of international model intercomparison projects such as CMIP and submissions to

IPCC. A key requirement is to make the intercomparisons as efficient as possible, while ensuring scientific credibility. To this aim, having efficient tools to find the data (stopping the data wrangling) and clear, transparent workflows are an important factor.

- Regional models: Regional climate models aim for downscaling climate fields produced by coarse resolution global models. Thereby, these regional models provide information at fine, grid scales more suitable for studies of regional phenomena and for application to vulnerability, impacts, and adaptation. Evaluation tools need to accommodate different grids, different temporal resolutions.

Those who participated in the study felt more thought needed to be given to how the climate impact community could usefully engage with ESM evaluation and how climate studies could in return benefit ESM evaluation (in terms of science and tools). Members of the community contacted during the study had no experience of using such tools themselves and stated they worked closely with climate modellers to advise on, for example, interpretation of model outputs.

GENREQ 2.0 Easy for any particular user to find/get what they want. Focus on the "user experience"

Developments of community tools should be user oriented. With the wide range of potential users, identified above, it will be difficult to accommodate all requirements, while at the same time making the tool usable by any particular user. As well as good documentation, training and support (see GENREQ 4.0 below), some users may not need the whole tool – just particular pre-processing functions, diagnostics, etc.

GENREQ 3.0: Efficient and easy to use: comparing set up time and run time of existing tools versus developing own tools

Stakeholders reported that they are often discouraged from using "off-the-shelf" tools, due to the investment in time to understand the tool and its requirements. Users said typically they were prepared to invest a few hours/~day for initial set up.

GENREQ 4.0: Good documentation, training and support

In order to support REQ 2.0 and 3.0 above good documentation, training and support are essential. Increasingly these should be provided online and include user mailing lists, help desks.
Careful organisation and search of documentation is essential to:

- answer "how to..." questions
- guide users through the range of data, diagnostics and metrics available
- keep them up to date

- support users in developing their own recipes, diagnostics, metrics or introducing new data sources
- inviting contributions from the community.

It was clear from the interviews that local support and interaction between tools developers and users was valued by those who have it, (e.g. at BSC, SMHI, DLR for ESMValTool) to deal with issues such as native formats, local IT environments, configurations, local security.

GENREQ 5.0: Transparent and traceable: no "black boxes", provenance of information easy to track

Scientific users are very reluctant to use off-the-shelf tools if they cannot see how the diagnostics and metrics are produced, following each step in detail. Provenance information is particularly important for outputs submitted to IPCC which has set high standards on openness, rigour, transparency and traceability.

GENREQ 6.0: Reliable, tested: certified

Linked to REQ 5.0, users need to know the level of testing that has been carried out, so that, for example, results are reproducible. Some users expressed an interest in a certification system. Details of how this would be achieved in practice need further thought.

GENREQ 7.0: Confident that it will be maintained and developed

Those users reluctant to use off-the-shelf tools needed assurance that a long term sustainable development strategy and related funding are in place. Many tools, such as ESMValTool, have relied to a significant degree on short term project funding, making it more difficult to set out a strategic vision or engage with a very wide community.

GENREQ 8.0: Solution to growing data volumes (becoming problematic for evaluation): High temporal and spatial resolution simulations

Climate data, including from international model intercomparison projects such as CMIP6 present many challenges in including capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and sampling. Ensembles, high temporal and spatial resolution simulations and newly integrated physical and biogeochemical processes are compounding the challenges. CMIP6 consists of simulations from approximately 100 climate models produced across 49 different modelling groups. Datasets include over 2000 variables. Adding observational data and reanalyses adds to the big data challenges. The available datasets are catalogued and held by the Earth System Grid Federation[5], a collaboration that develops, deploys and maintains software infrastructure for the management, dissemination, and analysis of model output and observational data. Development of evaluation tools needs to work closely with data platforms such as ESGF to ensure that users can get access to the data they need in an acceptable time and that the dataset provides good value to the user.

GENREQ 9.0 Interoperable with other tools – Python and other languages

It is recognised that users of ESM evaluation tools will need access to a wide range of other software packages to manage their work including for file storage, scheduling, workflow management, computation, data transfer. The wide range of possibilities, including the architecture of future international model intercomparison, makes this a challenging area.
Study stakeholders recognised the rapid evolution of IT, including computer languages. To facilitate interoperability with other tools there is high support for evaluation tools in Python. Overall, there was no consensus on the number of other languages to be supported. Software developers pointed to the difficulties of maintaining code in a number of languages, while others felt that pushing for a limited number of languages helped democratizing them.

GENREQ 10.0: GUIs (cf IS-ENES3 plans), APIs, click and play, toolbox – not just command line

To attract a wider range of users it was felt important to develop Graphical User Interfaces, APIs and toolboxes – not just command line interfaces.

GENREQ 11.0: Open source

Stakeholders supported open-source development of evaluation tools.

GENREQ 12.0: Diversity of tools

Users felt that while endorsement and development of a European community tool is welcomed, they also commented that there should be no "obligation" to use only one tool: diversity is needed for example, to stimulate innovation or to publish results from complex process studies where bespoke diagnostics and metrics needed to be developed.

[5] https://esgf.llnl.gov/

3.4 Future directions and scientific needs

SCIREQ 1.0: Ability to pick and choose and adapt

Given the vast range of climate science, stakeholders accepted that the scientific content of the evaluation tools (recipes, data access, pre-processing, diagnostics, metrics) could not, and arguably should not, be exhaustive, but rather tools should provide a rich collection which enables users to pick and satisfy their needs easily. In general, those interviewed accepted they may need to adapt or (re)write code to get the exact diagnostic/plot they need. Process oriented diagnostics are particularly difficult to standardise/automate. Users were also generally happy to contribute their code but cautioned against letting community tools become a “dumping ground” for everyone’s favourite diagnostic.

SCIREQ 2.0: Areas for development

Those consulted during the study pointed to various aspects of model evaluation, which in their opinion, needed further development. These are listed here, but not ranked or prioritised. More work needs to be done with the community to assess what is available (users are not always aware) and which diagnostics or performance metrics rank high, e.g. by demonstrating their importance in ESM evaluation by reference to peer-reviewed literature.

- Assessment metrics for model variability on various timescales (eg MJO, ENSO, NAO…)
- Analysis of extremes.
- Analysis of model biases.
- Which models – not just CMIP, eg Coupled Chemistry Intercomparison Project, more observational data.
- Evaluation of regional/local models, downscaling (not just global).
- High temporal and spatial resolution simulations (e.g. precipitation was mentioned by several interviewees where spatial and temporal granularity was important to advise impact communities).
- Catering for new/different grids.
- Machine learning and AI for ESM evaluation – interviewees were divided on the benefits of such techniques. For some, techniques associated with “big data analytics” were worth exploring to reduce the problems of large data volumes. Others thought the techniques offered no benefit for climate science.
- On the fly post-processing and diagnostics while the model is running (cf. developments at BSC). Some could foresee a time when only diagnostics and metrics were stored, rather than the vast arrays of model outputs but there was no consensus on this.
- Some interviewees were interested in looking at geographically distributed computing to reduce the burden of data transfer.

- On a related point, greater use of cloud computing for diagnostics and metrics could help to solve the data volume problem.
- Providing "standard" evaluations at the end of model runs to provide a mark of quality ("fitness for purpose"). This was a controversial point as a number of interviewees were adamant that evaluation should not be used to imply one model was "better" than another.
- Making tools usable by different communities, e.g. impacts, adaptation, paleoclimate…but with caution. More needs to be understood about the requirements of these communities and how they could use ESM evaluations in a scientifically robust manner. There was concern that the complexity of ESM diagnostics and metrics could easily mislead.
- Fast cycles of releases: 2-3 minor releases per year to keep evaluation software up to date.
- Improved user engagement, communications, training, tutorials, help desks, promoting success stories. At the most basic level many potential users of existing tools are not aware of their availability.

SCIREQ 3.0: Keep a close eye on CMIP7 developments and requirements

The cycles of CMIP experiments and inputs into IPCC have driven many of the developments of tools such as ESMValTool and PMP[6].
Stakeholders commented it was too early to say how future ESM evaluation will evolve internationally and the architecture and contents of CMIP7 are at an early stage discussion. Some commented that the growth is the scope of CMIP and the schedule of experiments is unsustainable and indeed may not be the best use of resources (do all centres need to do everything?) Changes in the CMIP system architecture (e.g. changes to distribute the ESGF infrastructure, less duplication), changes in development/analysis cycles and changes to number of experiments and who does what are likely to have an important impact on ESM evaluation needs.

[6] See for example the CMIP6 web site https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6

4. Feedback and implications for ESMValTool

4.1 Overview of ESMValTool

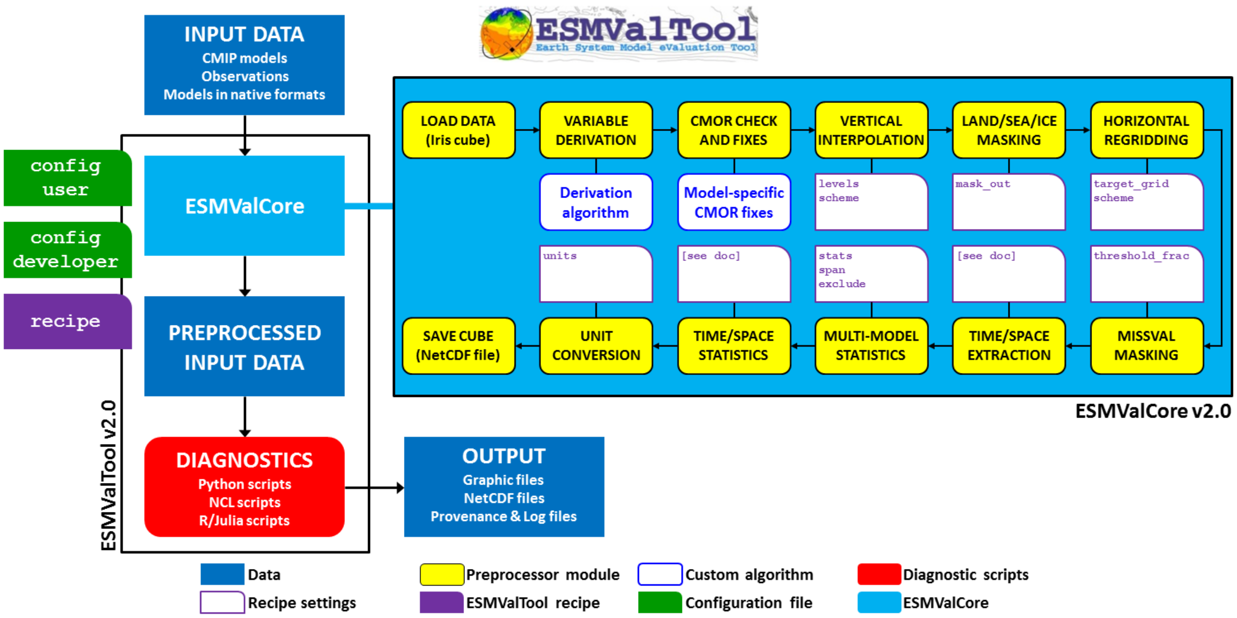

Figure 1:Schematic of the ESMValTool system architecture (https://docs.esmvaltool.org/en/latest/introduction.html#what-esmvaltool-can-do-for-you)

As noted throughout this report over the last decade significant progress has been made in model evaluation. For the CMIP community many baseline aspects of model evaluation can be performed efficiently to enable a systematic and rapid performance assessment of the large number of models using tools like ESMValTool and PMP.

Focusing on ESMValTool, it is available as open-source software on GitHub (https://github.com/ESMValGroup/ESMValTool) as well as pip, conda and docker container platform. The website at http://cmip-esmvaltool.dkrz.de/ shows results produced with the ESMValTool for CMIP5 and CMIP6 simulations.

Since the tool is freely available on GitHub, modelling groups participating in CMIP and others may make use of the package in other ways. For example, during the model development process in order to identify relative strengths and weaknesses of new model versions or they could run the tool locally before publishing the model output to the ESGF. ESMValTool is designed to work across ESGF nodes with the intent of ultimately expediting routine analysis by alleviating the needs for data transfer. The community is encouraged to consider contributing additional diagnostics and metrics.

4.2 Feedback from study participants

ESMValTool is held in high regard by study participants, even those who have until now taken different approaches to ESM evaluation. It has many of the attributes and addresses many of the requirements noted in Sections 3.3 and 3.4:

- "*I wish ESMValTool had been available, at least in its current form when I was doing my PhD – would have saved a lot of time – got results in half the time*." Study interviewee.
- "*Reduces data wrangling*." Study interviewee.
- Essentially it provides a workflow and an engine for performing computations within a framework – as shown in Figure 1. Interviewees particularly appreciated the data access and pre-processing functions of ESMValTool.
- Easily accessible - open-source software on Github.
- Results of using tool for CMIP5 and CMIP6 can be seen on DKRZ web site
    - 4884 results for CMIP5
    - 2065 results for CMIP6.
- Over 200 publications (Google Scholar) – it has high scientific credibility.
- Over 40 institutions are known to have an interest, but detailed user community is difficult to assess (anonymous use). Interviewees felt that the number of users was less important than the impact of the tool.
- Significant technical improvements in usability over the last 3 years, e.g. installation, supports multiple languages (Python, R, Julia, NCL), improved software engineering: testing, standards, version control, modular design.
- Can be run at ESGF centres (e.g. CEDA, DKRZ) or locally.
- Provides detailed information on provenance of results (input data and processing steps[7]).
- 80% of recipes run in ~20min, others take no more than half a day (user experience).
- Core development team at DLR praised for being very responsive.
- Training, documentation, help available on line.
- User mailing list, user engagement team in place and growing.
- Stops unnecessary duplication of evaluation tools across modelling/ evaluation groups.
- Easy to use and contribute.
- Ongoing active development, including as part of European research and development programmes, EU H2020 and Horizon Europe.

Interviewees confirmed that ESMValTool is a good candidate for continued development as a European ESM evaluation community tool with recognition that expertise and knowledge can be shared to accelerate climate science. While knowledge about ESMValTool is growing there still needs to be more systematic publicity and promotion. A number of those who are benefitting came across it by chance.

[7] https://is.enes.org/files/S42_ISENES3_GA1_ESMValTool_Broetz.pdf provides details of traceability and reproducibility in ESMValTool

4.3 Pointers on governance

Interviewees were reluctant to discuss governance in detail during the study, either because it was not of interest or because they did not want to prejudice ongoing discussions. Some thoughts on governance structures and organisational issues offered by interviewees are as follows in the context of a European community ESM evaluation tool:

- A European tool (as opposed to relying on tools from the USA for example) is important to maintain control and hence meet European priorities.
- The governance structure should avoid a single point of failure, i.e. the development and maintenance of the tool should rely on more than one institution.
- There are examples of good practice regarding governance, e.g. the ocean model NEMO (https://www.nemo-ocean.eu/ .
- Funding, outside short-term project funding is important: for long-term sustainability and to maintain a strategic vision, not unduly influenced by short term project priorities. One solution is long term, multi-institutional commitment and funding.
- The decision making process should be clear to users.
- The main governance body (of funders) should define the multi-year development strategy, keeping a distinction between the "Core" (Systems Team working to strict and agreed software standards) and the overall "Tool" with a wide range of diagnostics and metrics (closer to scientific developers). The Systems Team would ensure regular releases while a Science Team would advise on integration of new diagnostics and metrics based on an assessment of science needs. When the need arises, working groups could be created to gather the community expertise for advising on the development activities.

A key to continued successful development is getting the scientists and software developers to work together, at all levels, from day-to-day problem solving to strategic decision making.

- software developers need more understanding of the community needs
- scientists need to put more effort in using/understanding the tools.

5. Conclusions and recommendations

5.1 Conclusions

There was good support, amongst project participants, for the continued development of a European community ESM evaluation tool. The results are based on a qualitative survey using open-ended questions.

A diversity of approaches continues to stimulate innovation. Amongst these ESMValTool has emerged over recent years as a well regarded tool with many of the attributes and features required by a wide climate science community to carry out evaluations for model development and intercomparison and to support process studies. It is promoted by CMIP6 as a tool for model intercomparisons, including contributions to IPCC assessments.

The overarching approach which emerged from the survey is:

- Spend time identifying the main users.
- Make it easy for the users – focus on the "user experience", providing what the user needs/asks for.
- Provide a workflow and engine for outputting diagnostics and metrics in a common framework, removing the need to reinvent the wheel for commonly performed tasks.
- Users need to recognise that they will need to invest some time in familiarising themselves with their chosen tool.
- Distinguish between the "core" data access and pre-processing tasks which will be needed by many users and the overall tool. Specific diagnostics and metrics which may have a more limited user community. The core should be developed to strict software standards by a professional software team. There should be more flexibility in the development diagnostics and metrics including "user" contributions based on user needs.
- Ensure developments are open source.
- Provide long term governance, system administration, funding for developers.
- Avoid a single point of failure for governance, funding, and development.
- Organise ways to share knowledge and expertise efficiently and effectively, including bug fixes and more strategic decisions.
- Focus on "impact" rather than numbers of users.
- Be clear about the limitations of evaluation, e.g. when answering scientific and policy questions.

Areas of ongoing importance for ESMValTool (and other tools):

- Ensure excellent software engineering standards, particularly for common CORE tasks such as data access and pre-processing.
- Agree data standards and naming conventions for model and observational data.
- Reduce "data wrangling" as much as possible to increase time for productive science – this is a key attribute of successful evaluation tools.
- Provide transparency, provenance, and traceability of all steps.
- Enable easy intercomparisons between models, models and data. Include ways to avoid large data downloads to local sites.
- Ensure adequate testing of the tool and formalise regular releases.
- Provide comprehensive documentation, easy to navigate so that users can find what they want.
- Address licensing issues, including data inputs and code to ensure use of the tool is not restricted by restrictive licences.
- Make it flexible and easily extensible, but in a structured way

Based on user needs, support multiple coding languages (Python considered of high importance) and access methods (command line, GUI, API).

5.2 Recommendations

There is sufficient momentum behind ESMValTool to recommend its continued development as a European community tool for ESM evaluation and for encouraging engagement with a wide climate scientific community.

Going forward it will be important for data scientists, climate scientists and software developers to work closely together.

- software developers need more understanding of the community needs
- climate scientists need to put more effort in using/understanding the tools
- both climate scientists and software developers need to work with the data scientists who are making data available.

A clear long term strategy for the governance, funding and human resources is needed. We note that IS-ENES3 is already aware of these concerns and includes the ESMValTool in the sustainability strategy.

Appendix A: list of project participants

Study participants highlighted in green have direct hands-on knowledge of ESMValTool as developers and/or users.

Name	Affiliation	Project role	
		Study Steering Group	Interviewee
Pierre-Antoine Bretonniere	Barcelona Supercomputing Centre, Spain		√
Francisco J. Doblas-Reyes	Barcelona Supercomputing Centre, Spain		√
Martin Jury	Barcelona Supercomputing Centre, Spain		√
Kim Serradell	Barcelona Supercomputing Centre, Spain	√	
Javier Vegas	Barcelona Supercomputing Centre, Spain	√	√
Björn Brötz	DLR, Germany	√	√
Remi Kazeroni	DLR, Germany	√	
Stephan Kindermann	DKRZ, Germany		√
Olivier Boucher	Institut Pierre-Simon Laplace (IPSL), France		√
Eric Guilyardi	Institut Pierre-Simon Laplace (IPSL), France	√	√
Guillaume Levavasseur	Institut Pierre-Simon Laplace (IPSL), France		√
Sophie Morellon	Institut Pierre-Simon Laplace (IPSL), France	√	
Jérôme Servonnat	Institut Pierre-Simon Laplace (IPSL), France	√	√
Tom Crocker	Met Office, UK		√
Tamzin Palmer	Met Office, UK		√
Cath Senior	Met Office, UK		√
Fanny Adloff	National Centre for Atmospheric Science (NCAS), UK	√	√
Bryan Lawrence	National Centre for Atmospheric Science (NCAS) and University of Reading, UK		√
Andrew Gettelman	NCAR,USA		√
Klaus Zimmerman	SMHI, Sweden	√	√
Klaus Wyser	SMHI, Sweden		√
Martin Juckes	STFC, UK		√
Ranjini Swaminathan	University of Reading, UK		√

Appendix B: Interview guide

This interview guide was agreed with the Study Steering Group in July 2020. Some details of Annex A and Annex B of this interview guide may not be up to date.

Interview guide for ESM validation project

1. BACKGROUND

Earth System Model evaluation is becoming increasingly important and challenging. Model outputs are growing fast as spatial and temporal resolution increases and the number of ensemble runs per model increases. The range diagnostics needed is also growing significantly, as is the user community. Stakeholders and modelling groups in the IS-ENES 3 project (Infrastructure for the European Network for Earth System Modelling project running 2019 - 2022) are keen to be relevant to the growing crowd of model users who increasingly need some understanding of model evaluation. Therefore, as part of the new modelling infrastructure in IS-ENES 3, it has been decided to seek a better understanding of model evaluation needs. IS-ENES 3 seeks to provide an infrastructure that can be used as widely as possible. It is recognised that there are models and users with whom IS-ENES 3 has not so far engaged. The IS-ENES 3 project has commissioned Assimila Limited to undertake a wide reaching study of Earth System Model evaluation needs. The results will inform the development of evaluation tools in particular the ESMValTool (a community diagnostic and performance metrics tool for routine evaluation of Earth System Models in CMIP) or provide evidence why other approaches are needed.

GDPR, LEVEL OF DISCLOSURE AND ATTRIBUTION OF OPINIONS

For your information, your personal data (name, surname, affiliated institution, e-mail address) will be collected by Assimila Limited for the IS-ENES3 consortium. The purpose of the processing is the conduct of the IS-ENES3 Model Evaluation survey. The legal basis of the processing is Article 6 (1) lit. a General Data Protection Regulation (GDPR). Your data will be deleted at the end of the project (December 2022). In accordance with the European GDPR, you can ask for data rectification, data access and data erasure, or withdraw your consent by contacting Zof Stott (zof.stott[at]assimila.eu).

We will ask you to sign a consent form regarding storing of your personal data and levels of disclosure and attribution of opinions from the interview. The intention is to publish the report resulting from the consultation exercise and also to make interview summaries available to IS-

ENES 3. We recognise that you may wish to make some confidential, non-attributable comments. The consent form provides an opportunity for you to confirm the level of disclosure and attribution.

We will provide you an opportunity to check the transcripts of the interviews.

Are you happy for the interview to be recorded?

Question 1.1 Who are you? – describe of your role in your organisation/job title , particularly with respect to model evaluation and using ESM evaluation tools (eg model developer, ESM tool developer, data provider, model evaluator for XXXX purpose..).

Question 1.2 Which model evaluation tool(s) do you use, or contribute to (eg as developer, or data provider)

ESMValTool or other. If other please name and briefly describe, including use of own institution legacy scripts. Are there non-European, eg US or Japanese, evaluation tools you know about which we should consider in this study? If you do not currently use model evaluation tools briefly say why not and what your interest might be in the future.

---

2. CURRENT SITUATION

Question 2.1 How do you use it

To test a model, multiple models, against predecessor versions, observations.

How did you initially discover the tool(s) and how are you kept up to date with latest developments? How often do you use it, for how long? Ad hoc, routine use…Any comments to time, ease, power, system requirements.

How much do you rely on technical assistance to run the tools and workflows. How much interaction with the developers of the tool is needed and how do they handle communication and feedback.

Question 2.2 Why do you use it

For example: To develop the tool itself. To develop climate model/ESM. For model intercomparison purposes eg CMIP. To provide data for the community to use alongside the tools

To choose the "best model" for a particular scientific application, eg to get more insight into a particular scientific topic, eg tropical climate variability, monsoons, Southern Ocean processes, continental dry biases and soil hydrology-climate interactions, as well as atmospheric CO2 budgets, tropospheric and stratospheric ozone, and tropospheric aerosols…

To choose the "best model" for my commercial application…

Question 2.3 Which features are most important (list below comes mainly from ESMValTool)

- Facilitates the complex **evaluation of ESMs** and their simulations submitted to international Model Intercomparison Projects (e.g., **CMIP**).
- **Standardized model evaluation can be performed against observations**, against other models or to compare different versions of the same model.
- **Wide scope**: includes many diagnostics and performance metrics covering different aspects of the Earth System (dynamics, radiation, clouds, carbon cycle, chemistry, aerosol, sea-ice, etc.) and their interactions.
- **Well-established analysis**: standard namelists reproduce specific sets of diagnostics or performance metrics that have demonstrated their importance in ESM evaluation in the peer-reviewed literature.
- **Broad documentation**: user guide including creation date of running the script, version number, analyzed data (models and observations), applied diagnostics and variables, and corresponding references. This helps to increase the traceability and reproducibility of the results.
- **High flexibility**: **new diagnostics** and **more observational data** can be easily added.
- **Multi-language support**: Python, NCL, R... other open-source languages are possible.
- **CF/CMOR compliant**: data from many different projects can be handled (CMIP, obs4mips, ana4mips, CCMI, CCMVal, AEROCOM, etc.). Routines are provided to CMOR-ize non-compliant data.
- **Integration in modelling workflows**: for EMAC, NOAA-GFDL and NEMO, can be easily extended.

- **Extensively tested**: before release.
- **Quick technical assistance.**
- **Availability of training.**
- **Standardised outputs:** graphical formats, same colour bars..etc
- **Other**

Question 2.4 Which features do you use

Which recipes do you use to provide diagnostics/evaluation (see Appendix A for ESMValTool recipes, we shall not go through these in detail but broad areas of interest are relevant)

Do you write recipes/diagnostics. How do you interact? Via APIs? Command line? What is important, eg speed of familiarisation? Ease of use? Scope?

Question 2.5 What problems have you encountered using ESM evaluation tools?

Data – is it easy to find? Is new data easy to integrate? What about data formats – is this an issue (eg ESMValTool requires data in CMOR format – is this an hindrance?) Is the suite of recipes comprehensive? Are system requirements clearly defined? Documentation – is it adequate? Should ESM evaluation tools do more – for example?

---

3. FUTURE DEVELOPMENTS

(see development roadmap for ESMValTool in Annex B)

What is missing?

What would you like to see more of? (less of?)

ESMValTool has published a roadmap: Is the roadmap adequate – are there items missing?

Would standards help? For what?

4. LONG TERM STRATEGIC VISION FOR ESM VALIDATION TOOLS

Are multiple tools needed? Why? Should there be one, say CMIP approved, tool? Do you have views on flexibility/adaptability versus a more rigid framework? What/how should standards to developed? How should development (access, maintenance) be resourced? Do you have opinions about licensing? Any comments on governance or ESMValTool or evaluation tools in general?

Annex A ESMValTool Recipes (from ESMValTool web site July 2020)

Atmosphere

Blocking metrics and indices, teleconnections and weather regimes (MiLES)
Clouds
Cloud Regime Error Metric (CREM)
Combined Climate Extreme Index
Consecutive dry days
Evaluate water vapor short wave radiance absorption schemes of ESMs with the observations.
Diurnal temperature range
Extreme Events Indices (ETCCDI)
Diagnostics of stratospheric dynamics and chemistry
Heat wave and cold wave duration
Hydroclimatic intensity and extremes (HyInt)
Modes of variability
Precipitation quantile bias
Standardized Precipitation-Evapotranspiration Index (SPEI)
Drought characteristics following Martin (2018)
Stratosphere - Autoassess diagnostics
Stratosphere-troposphere coupling and annular modes indices (ZMNAM)
Thermodynamics of the Climate System - The Diagnostic Tool TheDiaTo v1.0
Zonal and Meridional Means

Climate metrics

Performance metrics for essential climate parameters
Single Model Performance Index (SMPI)

Future projections

Emergent constraints for equilibrium climate sensitivity
Emergent constraints on carbon cycle feedbacks
Emergent constraint on equilibrium climate sensitivity from global temperature variability
Emergent constraint on snow-albedo effect
Equilibrium climate sensitivity
Multiple ensemble diagnostic regression (MDER) for constraining future austral jet position
Transient Climate Response
Constraining future Indian Summer Monsoon projections with the present-day precipitation over the tropical western Pacific

IPCC
IPCC AR5 Chapter 9 (selected figures)
IPCC AR5 Chapter 12 (selected figures)

Land
Landcover - Albedo
Hydrological models - data pre-processing
Landcover diagnostics
Land and ocean components of the global carbon cycle
Runoff, Precipitation, Evapotranspiration

Ocean
Recipe for evaluating Arctic Ocean
Climate Variability Diagnostics Package (CVDP)
Nino indices, North Atlantic Oscillation (NAO), Southern Oscillation Index (SOI)
Ocean diagnostics
Ocean metrics

Other
Capacity factor of wind power: Ratio of average estimated power to theoretical maximum power
Ensemble Clustering - a cluster analysis tool for climate model simulations (EnsClus)
Multi-model products
RainFARM stochastic downscaling
Seaice feedback
Sea Ice
Seaice drift
Shapeselect
Toymodel

Annex B: ESMValTool development roadmap (from ESMValTool web site July 2020)

Time line	Project	Tasks
Q2 2020	CMIP6-DICAD	● Import capability of (native) EMAC2 output (M7)
Q4 2020	CRESCENDO	● Making use of observations from ESA CCI and obs4MIPs, extending the physical climate to include key biogeochemical diagnostics (LMU) ● New diagnostics for operational analysis of multi-model ESM projections (LMU)
Q3 2020-Q3 2021	ESA CCI CMUG Phase 3	Implementation of the following ESA CCI+ datasets including new diagnostics and metrics to exploit the data for model evaluation (D5.3): ● land surface temperature (MetO) ● long-lived ghg (CH4) (DLR) ● water vapor (DLR) ● sea surface salinity (BSC) ● ocean color (SMHI)
Q3 2020-Q4 2022	IS-ENES3	● Support for unstructured grids (D9.1, met.no) ● Quicklook system for online diagnostics (D9.2, MetO) ● ESGF coupling and distributed computation features (D9.3, UREAD-NCAS) ● Extension to evaluate regional climate models and different timescales (D9.4, BSC) ● IS-ENES3 ESMValTool version including coupling to community consensus developed packages such as the ENSO metrics package

		and technical improvements such as improved and enhanced automated testing (D9.5, DLR)
N/A	N/A	• Autoassess diagnostics full inclusion in ESMValTool
N/A	N/A	• ESMValTool as main monitoring tool for UKESM2 development and deployment
N/A	N/A	• Extending the public Python API for the full functionality of the preprocessor