Geosci. Model Dev. Discuss., 8, C4352–C4355, 2016
www.geosci-model-dev-discuss.net/8/C4352/2016/

Geoscientific Model Development Discussions

Interactive comment on "ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth System Models in CMIP" *by* V. Eyring et al.

V. Eyring et al.

veronika.eyring@dlr.de

Received and published: 8 April 2016

Reply to Anonymous Referee #1

We thank the reviewer for the helpful comments. We have now revised our manuscript in light of these and the other review comments we have received. A pointwise reply is given below.

In this very long paper the authors present a new diagnostic tool for comparing climate models against either observations or other models. The paper

is written very clearly and is easy to follow. As the ESMValTool is still under construction and is expected to add more functionalities in future I regard this paper as a snap-shot of the project. For me it's fine to publish it as is. I just have a few general comments/questions and one minor typo that I found.

General comments:

i) The ESMValTool is still in development. The single functions or namelists are explained in great detail. However, since this undertaking is evolving it would be nice to have some tool or platform to look for changes/additions to the existing namelists and descriptions of new functionalities. Is something like this planned?

Yes. The current version already includes a first implementation of Sphinx (http://www.sphinx-doc.org/en/stable/), which allows for an easier and automatic documentation method as the tool grows. In future releases, the ESMValTool code will be formatted to allow for automatic documentation using Sphinx. We added more details on code documentation using Sphinx as well as a reference to the "ESMValTool User's Guide" (i.e., the supplementary information) to section 2 of the revised manuscript.

ii) If new functions are build, is there a central place where the code is checked/reviewed or how is the quality of the tool being maintained?

Checking the tool quality is a responsibility of the core development team. For that, we implemented an automatic testing framework, which allows checking that every new development does not affect existing code. In term of code formatting, we followed the pep8 standard for Python, which we also adapted to check also NCL scripts. This is described in detail in the "ESMValTool User's Guide" (i.e., the supplementary information).

iii) The tool checks and corrects certain errors such as units and so on. But from experience there are ‘issues’ that are harder to detect, for example mistakes in sign conventions, soil moisture in Antarctica, zeros instead of missing values over land in the ocean files, . . . Mostly these problems are found after a while. So what I would like to say is that the know issues can be changed easily but what about the ones which are not expected/known? Are there any efforts to automatically search for inconsistencies?

The reformat routines are able to automatically spot errors in variable dimensionality, coordinates (names, ordering and units), variable units, missing values definition. Other less common errors in the data are hard to detect automatically, hence an automatic search method has not been implemented yet. However, errors in the data are usually evident once a diagnostic is applied. In such a case, users can take advantage of the fixing framework in the reformat routines and define project- and model-specific procedures to correct any kind of error in the input data.

v) I find it really helpful, that it can be used to compare a model with observations but also with other models or previous versions of a model. Hopefully, the latter results in more homogeneous data on the archives (see point iii).

Thanks for highlighting this feature. Indeed, the tool can be also applied to compare different versions/releases of a dataset. Modelling groups could apply the tool to check the quality of their data before submitting them.

Typo: pg 7584, line 6: ‘e.g. CMIP, models’ (i guess at least) should be ‘e.g., CMIP models’

For clarity, we have rephrased this sentence as follows: “against other models, e.g. CMIP5 models”.

Interactive comment on Geosci. Model Dev. Discuss., 8, 7541, 2015.

Geosci. Model Dev. Discuss., 8, C4356–C4360, 2016
www.geosci-model-dev-discuss.net/8/C4356/2016/

Geoscientific Model Development Discussions

Interactive comment on "ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth System Models in CMIP" *by* V. Eyring et al.

V. Eyring et al.

veronika.eyring@dlr.de

Received and published: 8 April 2016

Reply to Anonymous Referee #2

We thank the reviewer for the helpful comments. We have now revised our manuscript in light of these and the other review comments we have received. A pointwise reply is given below.

[0] Ditto what the first reviewer wrote.

[1] I like that the tool is focused and modularized on 'specific scientific themes'. Unfortunately for me, the subsequent descriptions are a bit tedious to read because each theme section seems to follow a pattern. That said: the paper does what it has to do! I have no substantive suggestions that would improve the pace of the presentation.

Thanks very much for supporting the modularized structure around specific scientific themes!

[2] Regridding is mentioned several times in the text and I assume that each module has used the appropriate interpolation method. For example (p7553), "Model output is linearly regridded". I assume this means bilinear interpolation. Most commonly, this method is used because it is fast and simple. However, being 'fast and simple' does not mean it is the most appropriate. In practice, if the variable being interpolated is smoothly varying, just about any interpolation method will produce reasonable results. However, bilinear interpolation may not be appropriate for variables that are fractal in space such 3-hrly and daily precipitation. I suggest that in each place where regridding is mentioned it should mention the type of interpolation used. This could simply be an adjective: (p7562) "After regridding all .." use "After bilinearly regridding all ..".

Agreed. We have revised the text making sure that the adopted regridding method is mentioned in each section (where appropriate).

[3] The text (p7589) states "One current limitation is the lack of parallelization." The most recent version of the NCAR CVDP (v4.0.0) has a Python driver that uses simple task parallelism to substantially reduce wall clock times. The driver uses standard Python functions (no custom functions). This approach

should be investigated for future use by the ESMValTool developers.

Thanks for this suggestion! We are currently planning to revise the parts of the code dealing with data preprocessing, which are the most time consuming operations in v1.0. The preprocessing includes common operation such as data reformatting and regridding. The goal is to move these operations to a higher level in the code structure, so that they can be performed in advance and in a parallel framework. This will be probably written in Python and the package you are suggesting could be useful.

[4] I note that there is wiki page (p7590) for developers and contributors. Like model development, developing data processing functionality is 'kinda' fun!!! The authors mention (p7548) a testing framework and code documentation. No details are mentioned. Sometimes developing good test codes can take more time than developing the processing function(s) they are testing. With regard to documentation, cryptic descriptions are better than nothing but *not* much better. I suggest encouraging (?requiring?), simple usage examples.

We aim at having a standardized code documentation based on Sphinx: the framework is already part of v1.0, but it has not been completely applied to existing code yet. Concerning automated testing, the developers are currently required to provide a test namelist together with their codes. The goal of such test namelists is to provide quick but yet comprehensive test cases and to serve as usage examples. Following the suggestion of the reviewer, we added more details on automated testing and on code documentation using Sphinx as well as a reference to the "ESMValTool User's Guide" (i.e., the supplementary information) to section 2 of the revised manuscript.

What is not mentioned at all? Ummm, let me think! Ah yes, now I remember:

USER SUPPORT. I am sure: (a) the tool's implementation and the components are perfect; (b) all users will carefully read the documentation; (c) all users will write clean, unambiguous structured code; and (d) all users will spend time trying to debug their codes. However, in the highly unlikely event that my assertions are not correct, how do users get support? To whom or what should questions be addressed? Should questions be sent to some central location? Will someone monitor the support location? Ultimately, who is responsible for user support?
Based upon experience, user support can be time consuming, tedious and frustrating. On the other hand, it can be rewarding. It can expose developers to different ways of thinking. It can offer insight into new development paths.

Following your suggestion we are setting up a user mailing list, where users can submit questions and ask for support. Once fully operational, the link to the mailing-list will be made available on the ESMValTool webpage at www.esmvaltool.org.

[5] Some journals have suggested that software tools should be referenced via a DOI or a link. Python, NCL and R are mentioned but there are no references to these tools.

- **The original R reference is the following. Ihaka and Gentleman are the original R developers. It is 20 years old but I could not find any better reference. Also, I could not find a specific R language DOI.**
  R: A Language for Data Analysis and Graphics Ross Ihaka and Robert Gentleman Journal of Computational and Graphical Statistics Vol. 5, No. 3 (Sep., 1996), pp. 299- 314 DOI: 10.2307/1390807.
- **Python: https://www.python.org/ I could not find a specific DOI. Perhaps this link is the best.**

- **Should NCL be spelled out in addition to the commonly used acronymn (NCL)? NCL (NCAR Command Language) NCL has a DOI. The NCL web page suggests the following citation:**
  The NCAR Command Language (Version 6.3.0) [Software]. (2016). Boulder, Colorado: UCAR/NCAR/CISL/TDD.
  http://dx.doi.org/10.5065/D6WD3XH5

We apologies for this omissions and agree these references should be added. Thanks for pointing us to the proper citations, which have been inserted in the manuscript. NCL has been spelled out as suggested.

I am happy to see that the ESMValTool will have a DOI!

We have assigned a DOI which is now given in the Code Availability Section.

Interactive comment on Geosci. Model Dev. Discuss., 8, 7541, 2015.

ESMValTool (v1.0) - A community diagnostic and performance metrics tool for routine evaluation of Earth System Models in CMIP

~~V.~~Veronika Eyring[1], ~~M.~~Mattia Righi[1], ~~M.~~Axel Lauer[1], Martin Evaldsson[2], ~~A. Lauer[1],~~ ~~S.~~Sabrina Wenzel[1], ~~C.~~Colin Jones[3,4], ~~A.~~Alessandro Anav[5], ~~O.~~Oliver Andrews[6], ~~I.~~Irene Cionni[7], ~~E.~~Edouard L. Davin[8], ~~C.~~Clara Deser[9], ~~C.~~Carsten Ehbrecht[10], ~~P.~~Pierre Friedlingstein[5], ~~P.~~Peter Gleckler[11], ~~K.-D.~~Klaus-Dirk Gottschaldt[1], ~~S.~~Stefan Hagemann[12], ~~M.~~Martin Juckes[13], ~~S.~~Stephan Kindermann[10], ~~J.~~John Krasting[14], ~~D.~~Dominik Kunert[1], ~~R.~~Richard Levine[4], ~~A.~~Alexander Loew[15,12], ~~J.~~Jarmo Mäkelä[16], ~~G.~~Gill Martin[4], ~~E.~~Erik Mason[14,17], ~~A.~~Adam Phillips[9], ~~S.~~Simon Read[18], ~~C.~~Catherine Rio[19], ~~R.~~Romain Roehrig[20], ~~D.~~Daniel Senftleben[1], ~~A.~~Andreas Sterl[21], ~~L.~~Lambertus H. van Ulft[21], ~~J.~~Jeremy Walton[4], Shiyu Wang[2], and ~~K.~~Keith D. Williams[4]

[1] Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany

[2] Swedish Meteorological and Hydrological Institute (SMHI), 60176 Norrköping, Sweden.

[3] University of Leeds, Leeds, UK

[4] Met Office Hadley Centre, Exeter, UK

[5] University of Exeter, Exeter, UK

[6] Tyndall Centre for Climate Change Research, School of Environmental Sciences, University of East Anglia, Norwich, UK

[7] Agenzia nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile (ENEA), Rome, Italy

[8] ETH Zurich, Switzerland

[9] National Center for Atmospheric Research (NCAR), Boulder, USA

[10] Deutsches Klimarechenzentrum, Hamburg, Germany

[11] Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, CA, USA

[12] Max-Planck-Institute for Meteorology, Hamburg, Germany

[13] National Centre for Atmospheric Science, British Atmospheric Data Centre, STFC Rutherford Appleton Laboratory, United Kingdom

[14] Geophysical Fluid Dynamics Laboratory/NOAA, Princeton, NJ, USA

[15] Ludwig~~-Maximilian University-~~Maximilians-Universität München, Munich, Germany

[16] Finnish Meteorological Institute, Finland

[17] Engility Corporation, Chantilly, VA, USA

[18] University of Reading, Reading, UK

[19] Institut Pierre Simon Laplace, Paris, France

~~[19] University of Reading, Reading, UK~~

[20] CNRM-GAME, Météo France and CNRS, Toulouse, France

[21] Royal Netherlands Meteorological Institute (KNMI), De Bilt, The Netherlands

Correspondence to: V. Eyring (veronika.eyring@dlr.de)

Abstract

A community diagnostics and performance metrics tool for the evaluation of Earth System Models (ESMs) has been developed that allows for routine comparison of single or multiple models, either against predecessor versions or against observations. The priority of the effort so far has been to target specific scientific themes focusing on selected Essential Climate Variables (ECVs), a range of known systematic biases common to ESMs, such as coupled tropical climate variability, monsoons, Southern Ocean processes, continental dry biases and soil hydrology-climate interactions, as well as atmospheric CO_2 budgets, tropospheric and stratospheric ozone, and tropospheric aerosols. The tool is being developed in such a way that additional analyses can easily be added. A set of standard namelists for each scientific topic reproduces specific sets of diagnostics

or performance metrics that have demonstrated their importance in ESM evaluation in the peer-reviewed literature. The Earth System Model Evaluation Tool (ESMValTool) is a community effort open to both users and developers encouraging open exchange of diagnostic source code and evaluation results from the CMIP ensemble. This will facilitate and improve ESM evaluation beyond the state-of-the-art and aims at supporting such activities within the Coupled Model Intercomparison Project (CMIP) and at individual modelling centres. Ultimately, we envisage running the ESMValTool alongside the Earth System Grid Federation (ESGF) as part of a more routine evaluation of CMIP model simulations while utilizing observations available in standard formats (obs4MIPs) or provided by the user.

1. Introduction

Earth System Model (ESM) evaluation with observations or reanalyses is performed both to understand the performance of a given model and to gauge the quality of a new model, either against predecessor versions or a wider set of models. Over the past decades, the benefits of multi-model intercomparison projects such as the Coupled Model Intercomparison Project (CMIP) have been demonstrated. Since the beginning of CMIP in 1995, participating models have been further developed, with more complex and higher resolution models joining in CMIP5 (Taylor et al., 2012) which supported the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5) ~~(IPCC, 2013).~~(IPCC, 2013). The main purpose of these internationally coordinated model experiments is to address outstanding scientific questions, to improve the understanding of climate, and to provide estimates of future climate change. Standardization of model output in a format that follows the Network Common Data Format (netCDF) Climate and Forecast (CF) Metadata Convention (http://cfconventions.org/) and collection of the model output on the Earth System Grid Federation (ESGF, http://esgf.llnl.gov/) facilitated multi-model analyses. However, CMIP has historically lacked a common analysis tool available that could operate directly on submitted model data and deliver a standard evaluation of models against observations.

An important new aspect ~~for CMIP6~~in the next phase of CMIP (i.e., CMIP6 (Eyring et al., 2015)) is a more distributed organization under the oversight of the CMIP Panel, where a set of standard model experiments, which were common across earlier CMIP cycles, the Diagnostic, Evaluation and Characterization of Klima (DECK) experiments and the CMIP6 ~~Historical Simulation~~historical simulations, will be used to broadly characterize model performance and sensitivity to standard external forcing. Standardization, coordination, common infrastructure, and documentation functions that make the simulation results and their main characteristics available to the broader community are envisaged to be a central part of CMIP6~~ (Meehl et al., 2014). The Earth System Model eValuation~~. The Earth System Model Evaluation Tool (ESMValTool) presented here is a community development that can be used as one of the documentation functions in CMIP to help diagnose and understand the origin and consequences of model biases and inter-model spread. Our goal is to develop an evaluation tool that users can run to produce well-established analyses of the CMIP models once the output becomes available on the ESGF. This is realized through text files that we refer to as standard namelists~~ that~~, each ~~call~~calling a certain set of diagnostics and performance metrics to reproduce analyses that have demonstrated to be of importance in ESM evaluation in previous peer-reviewed papers or assessment reports. Through this approach routine and systematic evaluation of model results can be made more efficient. The framework enables scientists to focus on developing more innovative analysis methods rather than constantly having to "re-invent the wheel". An additional purpose of the ESMValTool is to facilitate model evaluation at individual modelling centres, in particular to rapidly assess the performance of a new model against predecessor versions. Righi et al. (2015) and Jöckel et al. (2015) have applied a subset of the namelists presented here to evaluate a set of simulations using different configurations of the global ECHAM/MESSy Atmospheric Chemistry model (EMAC). In this paper we also highlight the integration of ESMValTool into modelling workflows – including models developed at NOAA's Geophysical Fluid Dynamics Laboratory (GFDL), the EMAC model, and the NEMO ocean model – through the use of the ESMValTool's reformatting routine capabilities.

In addition to standardized model output, the ESGF hosts observations for Model Intercomparison Projects (obs4MIPs~~, Teixeira et al. (2014))~~ (Ferraro et al., 2015; Teixeira et al., 2014)) and reanalyses data (ana4MIPs, https://www.earthsystemcog.org/projects/ana4mips). The ~~obs4MIP~~obs4MIPs and ~~ana4MIP~~ana4MIPs projects provide the community with access to CMIP-like data sets (in terms of variables, temporal and spatial frequencies, and time periods) of satellite data and reanalyses, together with the corresponding technical documentation. The ESMValTool makes use of these observations as well as observations available from other sources to evaluate the models. In several of the diagnostics and metrics, more than one observational data set or meteorological reanalysis is used to account for uncertainties in observations. This is crucial for assessing model performance in a more robust and scientifically valid way.

For the model evaluation we apply diagnostics and in several cases also performance metrics. Diagnostics (e.g., the calculation of zonal means or derived variables in comparison to observations) provide a qualitative comparison of the models with observations. Performance metrics are defined as a quantitative measure of agreement between a simulated and observed quantity which can be used to assess the performance of individual models or generation of models. Quantitative performance metrics are routinely calculated for numerical weather forecast models, but have been increasingly applied to Atmosphere-Ocean General Circulation Models (AOGCMs) or ESMs. Performance metrics used in these studies have mainly focused on climatological mean values of selected ECVs (Connolley and Bracegirdle, 2007; Gleckler et al., 2008; Pincus et al., 2008; Reichler and Kim, 2008), and only a few studies have developed process-based performance metrics (SPARC-CCMVal, 2010; Waugh and Eyring, 2008; Williams and Webb, 2009). The implementation of performance metrics in the ESMValTool enables a quantitative assessment of model improvements, both for different versions of individual ESMs and for different generations of model ensembles used in international assessments (e.g., CMIP5 versus CMIP6). Application of performance metrics to multiple models helps highlighting when and where one or ~~a few~~more models represent a particular process well. While quantitative metrics provide a valuable summary of overall model performance, they usually do not give information on how particular aspects of a model's simulation interact to determine the overall fidelity. For example, a model could simulate a mean state (and trend) in global mean surface temperature that agrees well with observations, but this could be due to compensating errors. To learn more about the sources of errors and uncertainties in models and thereby highlight specific areas requiring improvement, evaluation of the underlying processes and phenomena is necessary. A range of diagnostics and performance metrics focussing on a number of key processes are also included in ESMValTool.

This paper describes ESMValTool version 1.0 (v1.0) which is the first release of the tool to the wider community for application and further development as open source software. It demonstrates the use of the tool by showing example figures for each namelist for either all or a subset of CMIP5 models. Section 2 describes the technical aspects of the tool, and Section 3 the type of modelling and observational data currently supported by ESMValTool (v1.0). In Section 4 an overview of the namelists of ESMValTool (v1.0) is given along with their diagnostics and performance metrics and the variables and observations used. Section 5 describes the use of the ESMValTool in a typical model development cycle and evaluation workflow and Section 6 closes with a summary and an outlook.

2. Brief overview of the ESMValTool

In this section we give a brief overview of ESMValTool (v1.0) which is schematically depicted in Fig. 1. A detailed user's guide is provided in the ~~supplementary material~~Supplement.

The ESMValTool consists of a workflow manager and a number of diagnostic and graphical output scripts. It builds on a previously published diagnostic tool for chemistry-climate model evaluation (CCMVal-Diag Tool, (Gettelman et al. (2012))), but is different in its focus. In particular, it extends to ESMs by including diagnostics and performance metrics relevant for the coupled Earth system, and also focuses on evaluating models with a common set of diagnostics rather than being mostly flexible as the CCMVal-Diag tool. In addition, several technical and structural changes have been made that facilitate development by multiple users. The workflow manager is written in Python, while a multi-language support is provided in the diagnostic and the graphic routines. ~~The current version supports Python, NCL and R, but it can be extended to other open-source languages.~~The current version supports Python (www.python .org), the NCAR Command Language (NCL, 2016) and R (Ihaka and Gentleman, 1996), but it can be extended to other open-source languages. The ESMValTool is executed by invoking the *main.py* script, which takes a namelist as a single input argument. The namelists are text files written using the XML (eXtensible Markup Language) syntax and define the data to be read (models and observations), the variables to be analysed and the diagnostics to be applied. The XML-syntax has been chosen in order to allow users to express the relationship between these three elements (data, variables and diagnostics) in a structured, easy to use way.

Within the workflow, the input data are checked for compliance with the CF and Climate Model Output Rewriter (CMOR, http://pcmdi.github.io/cmor-site/tables.html) standards required by the tool (see Section 3) via a set of dedicated reformatting routines, which are also able to fix the most common errors in the input data (e.g., wrong coordinates, undefined or missing values, non-compliant units, etc.). It is additionally possible to define new variables using variable-specific scripts, for example ~~in order~~ to calculate the total column ozone from a 3D ozone field (tro3), temperature (ta) and surface pressure (~~psl~~ps). The diagnostic and graphic routines are written in a modular and flexible way so that they can be customized by the user via diagnostic-specific settings in the configuration file (cfg-file) and variable-specific settings (in the directory variable_~~def_dir)~~defs/) without changing the source code~~ of the workflow manager~~. These routines are complemented by a set of libraries, providing general-purpose code for the most common operations (statistical analyses, regridding tools, graphic styles, etc.). The output of the tool can be both ~~netCDF~~NetCDF and graphics files in various formats. In addition, a log file is written containing all the information of a specific call of the main script: creation date of running the script, version number, analysed data (models and observations), applied diagnostics and variables, and corresponding references. This helps to increase the traceability and reproducibility of the results.

To facilitate the development of new namelists and diagnostics by multiple developers from various institutions while preserving code quality and reliability, an automated testing framework is included in the package. This allows the developers to verify that modifications and new code are compatible with the existing code and do not change the results of existing diagnostics. Automated testing within the ESMValTool is implemented on two complementary levels:

- unittests are used to verify that small code units (e.g., functions/subroutines) provide the expected results.
- integration testing is used to verify that a diagnostic integrates well into the ESMValTool framework and that a diagnostic provides expected results. This is verified by comparison of the results against a set of reference data generated during the implementation of the diagnostic.

Each diagnostic is expected to produce a set of well-defined results, i.e. files in a variety of formats and types (e.g., graphics, data files, ASCII files). While testing results of a diagnostic, a special namelist file is executed by ESMValTool which runs a diagnostic on a limited set of test data only minimizing executing time for testing while ensuring that the diagnostic produces the correct results. The tests implemented include:

- file availability: a check that all required output data have been successfully generated by the diagnostic. A missing file is always an indicator for a failure of the program.
- file checksum: currently the MD5 checksum is used to verify that contents of a file are the same.
- graphics check: for graphic files an additional test is implemented which verifies that two graphical outputs are identical. This is in particular useful to verify that outputs of a diagnostic remain the same after code changes.

Unittests are implemented for each diagnostic independently using nose (https://nose.readthedocs.org/en/latest/). Test files are searched recursively, executed and a statistic on success and failures is provided at the end of the execution. In order to run integration tests for each diagnostic, a small script needs to be written once. As for the unittests, a summary of success and failures is provided as output (see the Supplement for details).

For the documentation of the code, Sphinx is used (http://sphinx-doc.org/~~).~~) to organize and format ESMValTool documentation, including text which has been extracted from source code. Sphinx can help to create documentation in a variety of formats, including HTML, LaTeX (and hence printable PDF), manual pages and plain text. Sphinx was originally developed for documenting Python code, and one of its features is that it is able – using the so-called autodoc extension – to extract documentation strings from Python source files and use them in the documentation it generates. This feature apparently does not exist for NCL source files (such as those which are used in ESMValTool), but it has been mimicked here via a Python script, which walks through a subset of the ESMValTool NCL scripts, extracts function names, argument lists and descriptions (from the comments immediately following the function definition), and assembles them in a subdirectory for usage with Sphinx. The documentation includes a listing of the functions, procedures, and plotting routines in order to encourage the reuse of existing code in multiple namelists.

3. Models and observations

The open-source release of ESMValTool (v1.0) that accompanies this paper is intended to work with CMIP5 model output, but the tool is compatible with any arbitrary model output, provided that it is in CF-compliant netCDF format and that the variables and metadata are following the CMOR tables and definitions. The namelists are designed such that it is straightforward to execute the same diagnostics with either CMIP DECK or CMIP6 model output rather than CMIP5 output, and these will be provided when the new simulations are available. As mentioned in the previous section, routines are provided for checking CF/CMOR compliance and fixing the most common minor flaws in the model output submitted to CMIP5. More substantial deviations from the required standards in the model output may be corrected via project- and model-specific procedures defined by the user and automatically applied within the workflow. The current reformatting routines are, however, not able to convert arbitrary model output to the full CF/CMOR standard. In this case, it is the responsibility of the individual modelling groups to perform that conversion. Currently, model-specific reformatting routines are provided for EMAC (Jöckel et al., 2015; Jöckel et al., 2010), the GFDL CM3 and ESM models (Donner et al., 2011; Dunne et al., 2012; Dunne et al., 2013), and for NEMO (Madec, 2008) which is the ocean model used in for example EC-Earth (Hazeleger et al., 2012). Users can develop similar reformatting routines specific to their model using the template included in the package allowing the tool to run directly on the original model output rather than having to reformat the model output to CF/CMOR beforehand.

The observations are organized in tiers. Where available, observations from the obs4MIPs and reanalysis from the ana4MIPs archives at the ESGF are used in the ESMValTool. These data sets form "Tier 1". Tier 1 data are freely available for download to be directly used by the tool since they are formatted following the CF/CMOR standard and do not need any additional processing. For other observational data sets, the user has to retrieve the data from their respective source and reformat them into the CF/CMOR standard. To facilitate this task, we provide specific reformatting routines for a large number of such data sets together with detailed information of the data source, as well as download and processing instructions (see Table 1). "Tier 2" includes other freely available data sets and "Tier 3" includes restricted data sets (e.g., requiring the user to accept a license agreement issued by the data owner). For Tier 2 and 3 data, links and help scripts are provided, so that these observations can be easily retrieved from their respective sources and processed by the user. A collection of all observational data used in ESMValTool (v1.0) is hosted at DLR and the ESGF nodes at BADC and DKRZ, but depending on the license terms of the observations these might not be publicly available.

4. Overview of namelists included in ESMValTool (v1.0)

A number of namelists have been included in ESMValTool (v1.0) that group a set of performance metrics and diagnostics for a given scientific topic. Namelists that focus on the evaluation of physical climate process for respectively, the atmosphere, ocean, and land surface are presented in Sections 4.1, 4.2, and 4.3. These can be applied to simulations with prescribed SSTs (i.e., AMIP runs) or the CMIP5 historical simulations (simulations for 1850 to present-day conducted with the best estimates of natural and anthropogenic climate forcing) that are run by either coupled AOGCMs or ESMs. Another set of namelists has been developed to evaluate biogeochemical biases present in ESMs when additional components of the Earth system such as the carbon cycle, atmospheric chemistry or aerosols are simulated interactively (Sections 4.4 and 4.5 for carbon cycle and aerosols/chemistry, respectively).

In each subsection, we first scientifically motivate the inclusion of the namelist by reviewing the main systematic biases in current ESMs and their importance and implications. We then give an overview of the namelists that can be used to evaluate such biases along with the diagnostics and performance metrics included, and the required variables and corresponding observations that are used in ESMValTool (v1.0). For each namelist we provide 1-2 example figures that are applied to either all or a subset of the CMIP5 models. An assessment of CMIP5 models is however not the focus of this paper. Rather, we attempt to illustrate how the namelists contained within ESMValTool (v1.0) can facilitate the development and evaluation of climate model performance in the targeted areas. Therefore, the results of each figure are only briefly described in each figure caption.

Table 1 provides a summary of all namelists included in ESMValTool (v1.0) along with information on the quantities and ESMValTool variable names for which the namelist is tested, the corresponding observations or reanalyses, the section and example figure in this paper, and references for the namelist. Table 2 then provides an overview of the diagnostics included for each namelist along with specific calculations, the plot type, settings in the configuration file (cfg-file), and comments.

4.1. Detection of systematic biases in the physical climate: atmosphere

4.1.1. Quantitative performance metrics for atmospheric ECVs

A starting point for the calculation of performance metrics is to assess the representation of simulated climatological mean states and the seasonal cycle for essential climate variables (ECVs, (GCOS (2010))). This is supported by a large observational effort to deliver long-term, high quality observations from different platforms and instruments (e.g., obs4MIPs and the ESA Climate Change Initiative (CCI)) and ongoing efforts to improve global reanalysis products (e.g., ana4MIPs).

Following (Gleckler et al. (2008)) and similar to Fig. 9.7 of (Flato et al. (2013)), a namelist has been implemented in the ESMValTool that produces a "portrait diagram" by calculating the relative space-time root-mean square error (RMSE) from the climatological mean seasonal cycle of historical simulations for selected variables [*namelist_perfmetrics_CMIP5.xml*]. In Fig. 2 the relative space-time RMSE for the CMIP5 historical simulations (1980-2005) against a reference observation and, where available, an ~~alternate~~alternative observational data set, is shown. ~~The code allows comparison of up to four observational data sets.~~ The overall mean bias can additionally be calculated and adding other statistical metrics~~ like the PDF Skill Score introduced in Section 4.4.1~~

is straightforward. Different normalizations (mean, median, centered median) can be chosen and the multi model mean/median can also be added. In order to calculate the RMSE, the data is regridded to a common grid using a bilinear interpolation method. The user can select which grid to use as a target grid. The results shown in this section have been obtained after regridding the data to the grid of the reference dataset. With this namelist it is also possible to perform more in-depth analyses of the ECVs, by calculating seasonal cycles, Taylor diagrams (Taylor, 2001), zonally averaged vertical profiles and latitude-longitude maps. In the latter two cases, it is also possible to produce difference plots between a given model and a reference (usually the observational data set) or between two versions of the same model, and to apply a statistical test to highlight significant differences. As an example, Fig. 3 (left panel) shows the zonal profile of seasonal mean temperature differences between the MPI-ESM-LR model (Giorgetta et al., 2013) and ERA-Interim reanalysis (Dee et al., 2011), and Fig. 3 (right panel) a Taylor diagram for temperature at 850 hPa for CMIP5 models compared to ERA-Interim. A similar analysis can be performed with *namelist_righi15gmd_ECVs.xml*, which reproduces the ECV plots of Righi et al. (2015) for a set of EMAC simulations.

Tested variables in ESMValTool (v1.0) that are shown is Fig. 2 are selected levels of temperature (ta), eastward (ua) and northward wind (va), geopotential height (zg), and specific humidity (hus), as well as near-surface air temperature (tas), precipitation (pr), all-sky longwave (rlut) and shortwave (rsut) radiation, long-wave (LW_CRE) and shortwave (SW_CRE) cloud radiative effect, and aerosol optical depth (AOD) at 550 nm (od550aer). The models are evaluated against a wide range of observations and reanalysis data: ERA-Interim and NCEP (Kistler et al., 2001) for temperature, winds and geopotential height, AIRS (Aumann et al., 2003) for specific humidity, CERES-EBAF for radiation (Wielicki et al., 1996), Global Precipitation Climatology Project (GPCP, Adler et al. (2003)) for precipitation, Moderate Resolution Imaging Spectrometer (MODIS, (Shi et al. (2011))) and the ESA CCI aerosol data (Kinne et al., 2015) for AOD. Additional observations or reanalyses can be provided by the user for these variables and easily added. The tool can also be applied to additional variables if the required observations are made available in an ESMValTool compatible format (see Section 2 and ~~supplementary material~~Supplement).

4.1.2. Multi-model mean bias for temperature and precipitation

Near-surface air temperature (tas) and precipitation (pr) are the two variables most commonly requested by users of ESM simulations. Often, diagnostics for tas and pr are shown for the multi-model mean of an ensemble. Both of these variables are the end result of numerous interacting processes in the models, making it challenging to understand and improve biases in these quantities. For example, near surface air temperature biases depend on the models' representation of radiation, convection, clouds, land characteristics, surface fluxes, as well as atmospheric circulation and turbulent transport ~~Flato et al. (2013),~~(Flato et al., 2013), each with their own potential biases that may either augment or oppose one another.

The *namelist_flato13ipcc.xml* reproduces a subset of the figures from the climate model evaluation chapter of IPCC AR5 (Chapter 9, (Flato et al. (2013))). This namelist will be further developed and a more complete version included in future releases. The diagnostic that calculates the multi-model mean bias compared to a reference data set is part of this namelist and reproduces Figures 9.2 and 9.4 of (Flato et al. (2013)). Figure 4 shows the CMIP5 multi-model average as absolute values and as biases relative to ERA-Interim and the GPCP data for the annual mean surface air temperature and precipitation, respectively. Model output is ~~linearly~~ regridded using bilinear interpolation to the reanalysis or observational grid by default, but alternative options that can be set in the cfg-file include regridding of the data to the lowest or highest resolution grid in the entire input data set. Such figures can also be produced for individual seasons as well as for a single model simulation or other 2D variables if suitable observations are provided.

4.1.3. Monsoon

Monsoon systems represent the dominant seasonal climate variation in the tropics, with profound socio-economic impacts. Current ESMs still struggle to capture the major features of both the South Asian summer monsoon (SASM, Section 4.1.3.1) and the West African monsoon (WAM, Section 4.1.3.2). Sperber et al. (2013) and Roehrig et al. (2013) provide comprehensive assessments of the ability of CMIP5 models to represent these two monsoon systems. By implementing diagnostics from these two studies into ESMValTool (v1.0), we aim to facilitate continuous monitoring of progress in simulating the SASM and WAM systems in ESMs.

4.1.3.1. South Asian summer monsoon (SASM)

While individual models vary in their simulations of the SASM, there are known biases in ESMs that span a range of temporal and spatial scales. The namelists in the ESMValTool are targeted toward analysing these biases in a systematic way. Climatological mean biases include excess precipitation over the equatorial Indian Ocean, too little precipitation over the Indian subcontinent and excess precipitation over orography such as the southern slopes of the Himalayas (Annamalai et al., 2007; Bollasina and Nigam, 2009; Sperber et al., 2013), see also Fig. 4. The monsoon onset is typically too late in the models, and the boreal summer intra-seasonal oscillation (BSISO), which has a particularly large socio-economic impact in South Asia, is often weak or not present (Sabeerali et al., 2013). Monsoon low pressure systems, which generate many of the most intense rain events during the monsoon (Krishnamurthy and Misra, 2011) are often too infrequent and weak (Stowasser et al., 2009). In coupled models, biases in SSTs, evaporation, precipitation and air-sea coupling are common (Bollasina and Nigam, 2009) and have been shown to affect both present-day simulations and future projections (Levine et al., 2013). Interannual teleconnections with ENSO (Lin et al., 2008) and the Indian Ocean Dipole (Ashok et al., 2004; Cherchi and Navarra, 2013) are also not well-captured (Turner et al., 2005).

Three SASM namelists for the basic climatology, seasonal cycle, intra-seasonal and inter-annual variability and key teleconnections have been implemented into the ESMValTool focusing on SASM rainfall and horizontal winds in June-September (JJAS) [*namelist_SAMonsoon.xml*, *namelist_SAMonsoon_AMIP.xml*, *namelist_SAMonsoon_daily.xml*]. Rainfall and wind climatologies, including their pattern correlations and RMSE against observations, are similar to the metrics proposed by the Climate Variability and Predictability (CLIVAR) Asian–Australian Monsoon Panel (AAMP) Diagnostics Task Team and used by Sperber et al. (2013). Diagnostics for determining global monsoon domains and intensity follow the definition of (Wang et al. (2012)) where the global precipitation intensity is calculated from the difference between the hemispheric summer (May-September in the Northern Hemisphere, November-March in the Southern Hemisphere) and winter (vice versa) mean values, and the global monsoon domain is defined by those areas where the precipitation intensity exceeds 2.0 mm/day and the summer precipitation is > 0.55 x the annual precipitation (Fig. 5). Seasonal cycle diagnostics include monthly rainfall over the Indian region (5°-30°N, 65°-95°E) and dynamical indices based on wind-shear (Goswami et al., 1999; Wang and Fan, 1999; Webster and Yang, 1992). Figure 6 shows examples of the seasonal cycle of area-averaged Indian rainfall from selected CMIP5 models and their AMIP counterparts. The namelists include diagnostics to calculate maps of inter-annual standard deviation of JJAS rainfall and horizontal winds at 850 hPa and 200 hPa, and maps of teleconnection diagnostics between Nino3.4 SSTs (defined by the region 190°-240°E, 5°S to 5°N) and JJAS precipitation across the monsoon region (30°S to 30°N, 40°-300°E) following (Sperber et al., 2013). To generate difference maps, data are first regridded using an area-conservative binning and using the lowest resolution grid as target. For atmosphere-only models, we also evaluate their ability to represent year to year monsoon variability directly against time-equivalent observations to ~~see if~~check whether models, given correct inter-annual SST forcing, can reproduce observed year to year variations and significant events occurring in particular years. This evaluation is done by plotting the time-series across specified years of standardized anomalies (normalized by climatology) of JJAS-averaged dynamical indices and area-averaged JJAS precipitation over the Indian region (defined above) for both the models and observations. Namelists for intra-seasonal variability include maps of standard deviation of 30-50 day filtered daily rainfall, with area-averaged values for key regions including the Bay of Bengal (10°-20°N, 80°-100°E) and the Eastern equatorial Indian Ocean (10°S-10°N, 80°-100°E) given in the plot titles. To illustrate the northward and eastward propagation of the BSISO, Hovmöller lag-longitude and lag-latitude diagrams show either the latitude-averaged (10ºS-10ºN) and plotted for 60º-160ºE, or longitude-averaged (80ºE-100ºE) and plotted for 10ºS-30ºN, anomalies of 30-80 day filtered daily rainfall correlated against intraseasonal precipitation at the Indian Ocean reference point (75ºE-100ºE, 10ºS-5ºN). These use a slightly modified (for season, region and filtering band) version of the existing Madden-Julian Oscillation (MJO) NCL scripts, available at https://www.ncl.ucar.edu/Applications/mjoclivar.shtml, that are based on the recommendations from the US CLIVAR MJO Working Group (Waliser et al., 2009) and are similar to those shown in Lin et al. (2008) and used in Section 4.1.4.2 for the MJO.

Tested variables in ESMValTool (v1.0), some of which are illustrated in Figs. 5 and 6, include precipitation (pr), eastward (ua) and northward wind (va) at various levels, and skin temperature (ts). The primary reference data sets are ERA-Interim for horizontal winds, Tropical Rainfall Measuring Mission 3B43 version 7 (TRMM-3B43-v7; Huffman et al. (2007) for rainfall and HadISST ~~(Rayner et al., 2003)~~(Rayner et al., 2003) for SST, although the models are evaluated against a wide range of other observational precipitation data sets (see Table 1) and an alternate ~~reanalyses~~reanalysis data set: the Modern-Era Retrospective Analysis for Research and Applications (MERRA; Rienecker et al. (2011)).

4.1.3.2. West African Monsoon Diagnostics

West Africa and the Sahel are highly dependent on seasonal rainfall associated with the WAM. Rainfall in the region exhibits strong inter-decadal variability (Nicholson et al., 2000), with major socio-economic impacts (Held et al., 2005). Projecting the future response of the WAM to increasing concentrations of greenhouse gases (GHG) is therefore of critical importance, as is the ability to make dependable forecasts of the WAM evolution on monthly to seasonal timescales. Current ESMs exhibit biases in their representation of both the mean state (Cook and Vizy, 2006; Roehrig et al., 2013) and temporal variability (Biasutti, 2013) of WAM. Such biases can affect the skill of monthly to seasonal predictions of the WAM as well as long term future projections. CMIP5 coupled models often exhibit warm SST biases in the equatorial Atlantic, which induce a southward shift of the WAM in summer (Richter et al., 2014). Because of the zonal symmetry, the 10°W-10°E meridional transect of any geophysical variable (see below) is particularly informative with respect to the main features of the WAM and their representation in climate models (Redelsperger et al., 2006). For instance, the JJAS-averaged Sahel rainfall has a large inter-model spread with biases ranging from +-50% of the observed value (Cook and Vizy, 2006; Roehrig et al., 2013). Differences in simulated surface air temperatures are large over the Sahel and Sahara, with deficiencies in the Saharan heat low inducing feedback errors on the WAM structure. Here, a correct simulation of the surface energy balance is critical, where biases related to the representation of clouds, aerosols and surface albedo (Roehrig et al., 2013). The seasonal cycle also shows large inter-model spread, pointing to deficiencies in the representation of key processes important for the seasonal dynamics of the WAM. Daily precipitation is highly intermittent over the Sahel, mainly caused by a few intense mesoscale convective systems during the monsoon season (Mathon et al., 2002). Intense mesoscale convective systems over Africa as well as the diurnal cycle of the WAM are still a challenge for most climate models (Roehrig et al., 2013). Improving the quality of the WAM in climate models is therefore urgently needed.

To evaluate key aspects of the WAM, two namelists have been implemented into ESMValTool (v1.0) [*namelist_WAMonsoon.xml, namelist_WAMonsoon_daily.xml]*]. These include maps and meridional transects (averages over 10°W to 10°E) that provide a climatological picture of the summer (JJAS) WAM structure: (i) precipitation (pr) for the mean position of the WAM, (ii) near-surface air temperature (tas) for biases in the Atlantic cold tongue and the Saharan heat low, (iii) horizontal winds (ua, va) for the mean position and intensity of the monsoon flow at 925 hPa and of the mid- (700 hPa) and upper-level (200 hPa) jets. The surface and top of the atmosphere (TOA) radiation budgets provide a picture of the radiative fluxes associated with the WAM. Figure 7 shows the meridional transect of summer-averaged precipitation over West Africa for a range of CMIP5 models as an example for this namelist. Diagnostic for the mean seasonal cycle of precipitation is also provided to evaluate the WAM onset and withdrawal. Finally, a set of diagnostics for the WAM intra-seasonal variability evaluates the ability of models to capture variability of precipitation on timescales associated with African easterly waves (3-10 day), the MJO (25-90 days) and more broadly the WAM intra-seasonal variability (1-90 days). The strong day-to-day intermittency of precipitation is also diagnosed using maps of 1-day autocorrelation of intra-seasonal precipitation anomalies (Roehrig et al., 2013). To perform the autocorrelation analysis, data is first regridded to a common 1°×1° map using a bilinear interpolation method, whereas for generating difference maps the same regridding method as for the SASM diagnostics is used (see Section 4.1.3.1). Observations for evaluation are based on the following data sets: GPCP version 2.2 and Tropical Rainfall Measuring Mission 3B43 version 7 (TRMM-3B43-v7, Huffman et al. (2007)) precipitation retrievals, Clouds and Earth's Radiant Energy Systems (CERES) Energy Balanced and Filled (EBAF) edition 2.6 radiation estimates (Loeb et al., 2009), NOAA daily TOA outgoing longwave radiation ~~(Liebmann and Smith, 1996),~~(Liebmann and Smith, 1996), ERA-Interim reanalysis for the dynamics.

4.1.4. Natural modes of climate variability

4.1.4.1. NCAR Climate Variability Diagnostics Package

Modes of natural climate variability from interannual to multi-decadal time scales are important as they have large impacts on regional and even global climate with attendant socio-economic impacts. Characterization of internal (i.e., unforced) climate variability is also important for the detection and attribution of externally-forced climate change signals (Deser et al., 2012; Deser et al., 2014). Internally-generated modes of variability also complicate model evaluation and intercomparison. As these modes are spontaneously generated, they do not need ~~not~~to exhibit the same chronological sequence in models as in nature. However, their statistical properties (e.g., time scale, autocorrelation, spectral characteristics, and spatial patterns) are captured to varying degrees of skill among climate models. Despite their importance, systematic evaluation of these modes remains a daunting task given the wide range to consider, the length of the data record needed to adequately characterize them, the importance of sub-surface oceanic processes and uncertainties in the observational records (Deser et al., 2010).

In order to assess natural modes of climate variability in models, the NCAR Climate Variability Diagnostics Package (CVDP) (Phillips et al., 2014) has been implemented into the ESMValTool. The CVDP has been developed as a standalone tool. To allow for easy updating of the CVDP once a new version is released, the structure of the CVDP is kept in its original form and a single namelist [*namelist_CVDP.xml*] has been written to enable the CVDP to be run directly within ESMValTool. The CVDP facilitates evaluation of the major modes of climate variability, including ENSO (Deser et al., 2010), PDO (Deser et al., 2010; Mantua et al., 1997), the Atlantic Multi-decadal Oscillation (AMO, Trenberth and Shea (2006)), the Atlantic Meridional Overturning

Circulation (AMOC, Danabasoglu et al. (2012)), and atmospheric teleconnection patterns such as the Northern and Southern Annular Modes (NAM (Hurrell and Deser, 2009; Thompson and Wallace, 2000) and SAM (Thompson and Wallace, 2000), respectively), North Atlantic Oscillation (NAO, Hurrell and Deser (2009)), and Pacific North and South American (PNA and PSA, respectively (Thompson and Wallace, 2000)) patterns. For details on the actual calculation of these modes in CVDP we refer to the original CVDP package and explanations available at http://www2.cesm.ucar.edu/working-groups/cvcwg/cvdp.

Depending on the climate mode analyzed, the CVDP package uses the following variables: precipitation (pr), sea level pressure (psl), near-surface air temperature (tas), skin temperature (ts), snow depth (snd), and basin-average ocean meridional overturning mass ~~streamfunction~~stream function (msftmyz). The models are evaluated against a wide range of observations and reanalysis data, for example NCEP for near-surface air temperature, HadISST for skin temperature, and the NOAA-CIRES Twentieth Century Reanalysis Project (Compo et al., 2011) for sea level pressure. Additional observations or reanalysis can be added by the user for these variables. The ESMValTool (v1.0) namelist runs on all CMIP5 models. As an example, Fig. 8 shows the representation of the PDO as simulated by 41 CMIP5 models and observations (HadISST) and Fig. 9 the mean AMOC from ~~15~~13 CMIP5 models.

4.1.4.2. Madden-Julian oscillation (MJO)

The MJO is the dominant mode of tropical intraseasonal variability (30-80 day) and has wide impacts on numerous regional climate and weather phenomena (Madden and Julian, 1971). Associated with enhanced convection in the tropics, the MJO exerts a significant influence on monsoon precipitation, e.g. on the South Asian Monsoon (Pai et al., 2011) and on the west African monsoon (Alaka and Maloney, 2012). The eastward propagation of the MJO into the West Pacific can trigger the onset of some El Nino events (Feng et al., 2015; Hoell et al., 2014). The MJO also influences tropical cyclogenesis in various ocean basins (Klotzbach, 2014). Increased vertical resolution in the atmosphere and better and representation of stratospheric processes have led to an improvement in MJO fidelity in CMIP5 compared with CMIP3 (Lin et al., 2006). However, current generation models still struggle to adequately capture the eastward propagation of the MJO (Hung et al., 2013) and the variance intensity is typically too weak. Identifying and reducing such biases will be important for ESMs to accurately represent important climate phenomena, such as regional precipitation variability in the tropics arising through the differing impact of MJO phases on ENSO and ENSO forced regional climate anomalies (Hoell et al., 2014).

To assess the main MJO features in ESMs, a namelist with a number of diagnostics developed by the US CLIVAR MJO Working Group (Kim et al., 2009; Waliser et al., 2009) has been implemented in the ESMValTool (v1.0) [*namelist_mjo_mean_state.xml, namelist_mjo_daily.xml*]. These diagnostics are calculated using precipitation (pr), outgoing longwave radiation (OLR) (rlut), eastward (ua) and northward wind (va) at 850 hPa (u850) and 200 hPa (u200) against various observations and reanalysis data sets for boreal summer (May-October) and winter (November-April).

Observation and reanalysis data sets include GPCP-1DD for precipitation, ERA-Interim and NCEP-DOE reanalysis 2 for wind components (Kanamitsu et al., 2002) and NOAA polar-orbiting satellite data for OLR (Liebmann and Smith, 1996). The majority of the scripts are based on example scripts at http://ncl.ucar.edu/Applications/mjoclivar.shtml. Daily data is required for most of the scripts. The basic diagnostics include mean seasonal state and 20-100 day bandpass filtered variance for precipitation and u850 in summer and winter. To better assess and understand model biases in the MJO, a number of more sophisticated diagnostics have also been implemented. These include; univariate empirical orthogonal function (EOF) analysis for 20-100 day bandpass filtered daily anomalies of precipitation, OLR, u850 and u200. To illustrate the northward and eastward propagation of the MJO, lag-longitude and lag-latitude diagrams show either the equatorial (latitude) averaged (10ºS-10ºN) or zonal (longitude) averaged (80ºE-100ºE) intraseasonal precipitation anomalies and u850 anomalies correlated against intraseasonal precipitation at the Indian Ocean reference point (75ºE-100ºE, 10ºS-5ºN). Similar figures can also be produced for other key variables and regions following the definitions of (Waliser et al. (2009)). To further explore the MJO intraseasonal variability, the wavenumber-frequency spectra for each season is calculated for individual variables. In addition, we also produce cross-spectral plots to quantify the coherence and phase relationships between precipitation and ~~U850~~u850. Figure 10 shows examples of boreal summer (May-October) wavenumber-frequency spectra of 10ºS-10ºN averaged daily precipitation from GPCP-1DD, ~~HadGEM~~HadGEM2-ES, MPI-ESM-LR and EC-~~EARTH~~Earth. Finally, we also calculate the multivariate combined EOF (CEOF) modes using equatorial averaged (15ºS-15ºN) daily anomalies of U850, U200 and OLR. This analysis demonstrates the relationship between lower- and upper-tropospheric wind anomalies and convection. To further illustrate the spatial-temporal structure of the MJO, the first two leading CEOFs are used to derive a composite MJO life cycle which highlights intraseasonal variability and northward/eastward propagation of the MJO. The data used in these diagnostics are regridded to a common 0.5°×0.5° grid using an area-conservative method.

4.1.5. Diurnal cycle

In addition to the previously discussed biases in precipitation, many ESMs that rely on parameterized convection exhibit biases related to the diurnal cycle and timing of precipitation. Over land, ESMs tend to simulate a diurnal cycle of continental convective precipitation in phase with insolation, while observed precipitation peaks in the early evening. This constitutes one of the endemic biases of ESMs, in which convective precipitation intensity is often related to atmospheric instability. This bias can have important implications for the simulated climate, as the timing of precipitation influences subsequent surface evaporation, and convective clouds affect radiation differently around noon or in late afternoon. The biases in the diurnal cycle are most pronounced over land areas and the diurnal cycles of convection and clouds during the day contribute to the continental warm bias ~~(Cheruy et al., 2014).~~(Cheruy et al., 2014). Similarly, biases in the diurnal cycle also exist over the ocean (Jiang et al., 2015). Another motivation for looking at the diurnal cycle in models is that its representation is more closely linked to the parameterizations of surface fluxes, boundary-layer, convection and cloud processes than any other diagnostics. The phase of precipitation and radiative fluxes during the day is the consequence of surface warming, boundary-layer turbulence mixing and cumulus clouds moistening, as well as of the triggering criteria used to activate deep convection, and the closure used to compute convective intensity. The evaluation of the diurnal cycle thus provides a direct insight into the representation of physical processes in a model. Recent efforts to improve the representation of the diurnal cycle of precipitation models include modifying the convective entrainment rate, revisiting the quasi-equilibrium hypothesis for shallow and deep convection, and adding a representation of key missing processes such as boundary-layer thermals or cold pools. We envisage that ESMValTool will help to quantify the impact of those improvements in the next generation of ESMs.

To help document progress made in the representation of the diurnal cycle of precipitation (pr) in models, a set of diagnostics has been implemented in ESMValTool. After regridding all data on a common 2.5°×2.5° grid using bilinear interpolation, the mean diurnal cycle computed every 3 hours is approximated at each grid-point by a sum of sine and cosine functions (first harmonic analysis) allowing to derive global maps of the amplitude and phase of maximum rainfall over the day. Mean diurnal cycle of precipitation is also provided over specific regions in the tropics. Over land, we contrast semi-arid (Sahel) and humid (Amazonia) regions as well as West-Africa and India. Over the ocean, we focus on the Gulf of Guinea, the Indian Ocean and the East and West Equatorial Pacific. We use TRMM 3B42 ~~V6~~V7, as a reference (~~http://mirador.gsfc.nasa.gov/collections/TRMM_3B42_daily__006.shtml~~http://mirador.gsfc.nasa.gov/collections/TRMM_3B42_daily__007.shtml). The ESMValTool also includes diagnostics for the evaluation of the diurnal cycle of radiative fluxes at the top of the atmosphere and at the surface, and their decomposition into LW and SW, total and clear~~-~~ sky components, however not all are available for all models from the CMIP5 archive. As a reference, we use 3-hourly SYN1deg CERES products (Wielicki et al., 1996), derived from measurements at top of the atmosphere and computed using a radiative transfer model at the surface (http://ceres.larc.nasa.gov/products.php?product=SYN1deg). These diagnostics provide a first insight into the representation of the diurnal cycle, but further analysis is required to understand the links between the model's parameterizations and the representation of the diurnal cycle, as well as the impact of errors in the diurnal cycle on other, slower timescale climate processes. Figure 11 shows the evaluation against TRMM observations of the mean diurnal cycle averaged over specific regions in the tropics for five summers (2004-2008) simulated by four CMIP5 ESMs.

4.1.6. Clouds

4.1.6.1. Clouds and radiation

Clouds are a key component of the climate system because of their large impact on the radiation budget as well as their crucial role in the hydrological cycle. The simulation of clouds in climate models has been challenging because of the many nonlinear processes involved (Boucher et al., 2013). Simulations of long-term mean cloud properties from CMIP3 and CMIP5 models show large biases compared ~~with~~to observations (Chen et al., 2011; Klein et al., 2013; Lauer and Hamilton, 2013). Such biases have a range of implications as they affect application of these models to investigate chemistry-climate interactions and aerosol-cloud interactions, while also having an impact on the climate sensitivity of the model.

The namelist *namelist_lauer13jclim.xml* computes the climatology and interannual variability of climate relevant cloud variables such as cloud radiative forcing, liquid and ice water path, and cloud cover and reproduces the evaluation results of Lauer and Hamilton (2013). The standard namelist includes a comparison of the geographical distribution of multi-year average cloud parameters from individual models and the multi-model mean with satellite observations. Taylor diagrams are generated that show the multi-year annual or seasonal average performance of individual models and the multi-model mean in reproducing satellite observations. The diagnostic routine also facilitates the assessment of the bias of the multi-model mean and zonal averages of individual models compared with satellite observations. Interannual variability is estimated as the relative temporal standard deviation from multi-year timeseries of data with the temporal standard deviations calculated from monthly anomalies after subtracting the climatological mean seasonal cycle. Data regridding is applied using a bilinear interpolation method and choosing the grid of the reference dataset as target. As an example, Fig. 12 shows the bias of the 20-year average (~~1986~~1985-2005) annual mean cloud radiative effects from CMIP5 models (multi-model mean) against the CERES EBAF satellite climatology (2001-2012) (Loeb et al., 2012; Loeb et al., 2009), similar to Flato et al. (2013) their Figure 9.5.

The cloud namelist focuses on precipitation (pr) and four cloud parameters that largely determine the impact of clouds on the radiation budget and thus climate in the model simulations: total cloud amount (clt), liquid water path (lwp), ice water path (iwp), and ~~ToA~~TOA cloud radiative effect (CRE) consisting of the longwave CRE and shortwave CRE that can also separately be evaluated with the performance metrics namelist (see Section 4.1.1). Precipitation is evaluated with GPCP data, total cloud amount with MODIS, liquid water path with passive-microwave satellite observations from the University of Wisconsin (O'Dell et al., 2008), and the ice water path with MODIS Cloud Model Intercomparison Project (MODIS-CFMIP, (Pincus et al. (2012)), (King et al. (2003))) data.

4.1.6.2. Quantitative performance assessment of cloud regimes

The cloud-climate radiative feedback process remains one of the largest sources of uncertainty in determining the climate sensitivity of models (Boucher et al., 2013). Traditionally, clouds have been evaluated in terms of their impact on the mean top of atmosphere fluxes. However, it is possible to achieve good performance on these quantities through compensating errors, for example boundary layer clouds may be too reflective but have insufficient horizontal coverage (Nam et al., 2012). Williams and Webb (2009) proposed a Cloud Regime Error Metric (CREM) which critically tests the ability of a model to simulate both the relative frequency of occurrence and the radiative properties correctly for a set of cloud regimes determined by the daily mean cloud top pressure, in-cloud albedo and fractional coverage at each grid-box. Having previously identified the regimes by clustering joint cloud-top pressure-optical depth histograms from the International Satellite Cloud Climatology Project (ISCCP, Rossow and Schiffer (1999)) as per (Williams and Webb (2009)), each daily model grid box is assigned to the regime cluster centroid with the closest cloud top pressure, in-cloud albedo and fractional coverage as determined by the 3-element Euclidean distance. The fraction of grid points assigned to each of the regimes and the mean radiative properties of those grid points are then compared to the observed values. This routine also uses a bilinear regridding method with a 2.5°×2.5° target grid.

This metric is now implemented in ESMValTool (v1.0), with references in the code to tables in the (Williams and Webb (2009)) study defining the cluster centroids [*namelist_williams09climdyn_CREM.xml*]. Required are daily data from ISCCP mean cloud albedo (albisccp), ISCCP Mean Cloud Top Pressure (pctisccp), ISCCP Total Total Cloud Fraction (cltisccp), TOA outgoing short- and long-wave radiation (rsut, rlut), TOA outgoing shortwave radiation (rlutcs), surface snow area fraction (snc) or surface snow amount (snw), and sea ice area fraction (sic). The metric has been applied over the period January ~~1979~~1985 to December ~~1983~~1987 to those CMIP5 models ~~that submitted~~with the required diagnostics (daily data) available for their AMIP simulation (see caption of Fig. 13). A perfect score with respect to ISCCP would be zero. (Williams and Webb (2009)) also compared data from the MODIS and the Earth Radiation Budget Experiment (ERBE, Barkstrom (1984)) to ISCCP in order to provide an estimate of observational uncertainty. This observational regime characteristic was found to be 0.96 as marked on Fig. 13 when calculated over the period March 1985 to February 1990. Hence a model with a score that is similar to this value can be considered to be within observational uncertainty, although it should be noted that this does not necessarily mean that the model lies within the observations for each regime. Error bars are not plotted since experience has shown that the metric has little sensitivity to interannual variability and models that are visibly different on Fig. 13 are likely to be significantly so. A minimum of two years, and ideally five years or more, of daily data are required for the scientific analysis.

4.2. Detection of systematic biases in the physical climate: ocean

4.2.1. Handling of ocean grids

Analysis of ocean model data from ESMs poses several unique challenges for analysis. First, in order to avoid numerical singularities in their calculations, ocean models often use irregular grids where the poles have been rotated or moved to be located over land areas. For example, the global configuration of the Nucleus for European Modelling of the Ocean (NEMO) framework uses a tripolar grid (Madec, 2008), with the three poles located over Siberia, Canada and Antarctica. Second, transports of scalar quantities (e.g., overturning ~~streamfunctions~~stream functions and heat transports) can only be calculated accurately on the original model grids as interpolation to other grids introduces errors. This means that, e.g. for the calculation of water transport through a strait, both the horizontal and vertical extent of the grids on which the u and v currents are defined is required. Therefore, this type of diagnostic can only be used for models for which all native grid information is available. State variables like SSTs, sea ice and salinity are regridded using grid information (i.e., coordinates, bounds, and cell areas) available in the ocean input files of the CMIP5 models. To create difference plots against observations or other models all data are regridded to a common grid (e.g., 1~~°x1~~°×1°) using the regridding functionality of the Earth System Modeling Framework (ESMF, https://www.ncl.ucar.edu/Applications/ESMF.shtml).

4.2.2. Southern Ocean Diagnostics

4.2.2.1. Southern Ocean mixed layer dynamics and surface turbulent fluxes

Earth system models often show large biases in the Southern Ocean mixed layer. For example, Sterl et al. (2012) showed that in EC-Earth/NEMO the Southern Ocean is too warm and salinity too low, while the mixed-layer is too shallow. These biases are not specific to EC-Earth, but are rather widespread. At the same time, values for Antarctic Circumpolar Current (ACC) transport vary between 90 and 264 Sv in CMIP5 models, with a mean of 155±51 Sv. The differences are associated with differences in the ACC density structure.

A namelist has been implemented in the ESMValTool to analyse these biases [*namelist_SouthernOcean.xml*]. With these diagnostics polar stereographic (difference) maps can be produced to compare monthly/annual mean model fields with corresponding ERA-Interim data. The patch recovery technique is applied to regrid data to a common 1°×1° grid. There are also scripts to plot the differences in the area mean vertical profiles of ocean temperature and salinity between models and data from the World Ocean Atlas (Antonov et al., 2010; Locarnini et al., 2010). The ocean mixed layer thickness from models can be compared with that obtained from the Argo floats (Dong et al., 2008). Finally, the ACC strength, as measured by water mass transport through the Drake Passage, is calculated using the same method as in the CDFTOOLS package (CDFTOOLS, http://servforge.legi.grenoble-inp.fr/projects/CDFTOOL). This diagnostic can be used to calculate the transport through other sections as well, but is presently only available for NEMO/ORCA1 output, for which all grid information is available. The required variables for the comparison with ERA-Interim are sea surface temperature (tos), downward heat flux (hfds, calculated from ERA-Interim by summing the surface latent and sensible heat flux and the net shortwave and longwave fluxes (hfls+hfss+rsns+rlns)), water flux (wfpe, calculated by summing precipitation and evaporation (pr+evspsbl)) and the wind stress components (tauu and tauv). For the comparison with the World Ocean Atlas 2009 data (WOA09) sea surface salinity (sos), sea water salinity (so) and temperature (to) are required variables. For the comparison with the Argo floats the ocean mixed layer thickness (mlotst) is required. Finally the two components of sea water velocity (uo and vo) are required for the volume transport calculation. Some example figures from this set of diagnostic scripts are shown for EC-Earth in Fig. 14.

4.2.2.2. Atmospheric processes forcing the Southern Ocean

One leading cause of SST biases in the Southern Ocean is systematic biases in surface radiation fluxes (Trenberth and Fasullo, 2010) coupled with systematic errors in macrophysical (e.g. cloud amount) and microphysical (e.g. frequency of mixed-phase clouds) cloud properties (Bodas-Salcedo et al., 2014).

A namelist has been implemented into the ESMValTool that compares model estimates of cloud, radiation and surface turbulent flux variables over the Southern Ocean with suitable observations [*namelist_SouthernHemisphere.xml*]. Due to the lack of surface/in-situ observations over the Southern Ocean, remotely sensed data can be subject to considerable uncertainty (Mace, 2010). While this ~~is~~ uncertainty is not explicitly addressed in ESMValTool (v1.0), in future releases we will include a number of alternative satellite based data sets for cloud variables (e.g., MISR, MODIS, ISCCP) as well as new methods under development to derive surface turbulent flux estimates constrained by observed TOA radiation flux estimates and atmospheric energy divergence derived from reanalysis products (Trenberth and Fasullo, 2008). Inclusion of multiple satellite-based estimates will provide some estimate of observational uncertainty over the region. Variables analysed include (i) total cloud cover (clt), vertically integrated cloud liquid water and cloud ice water (clwvi, clivi) (ii) surface/ (TOA) downward/outgoing total sky and clear~~-~~ sky short wave and longwave radiation fluxes (rsds, rsdcs, rlds, rldscs / rsut, rsutcs, rlut, rlutcs) and (iii) surface turbulent latent and sensible heat fluxes (hfls, hfss). Observational constraints are derived from, respectively; cloud: CloudSat level 3 data (Stephens et al., 2002), radiation: CERES-EBAF level 3 Ed2 data and surface turbulent fluxes: WHOI-OAflux (Yu et al., 2008).

The following diagnostics are calculated with accompanying plots: (i) Seasonal mean absolute-value and difference maps for model data versus observations covering the Southern Ocean region (30°S-65°S) for all variables. (ii) Mean seasonal cycles using zonal means averaged separately over three latitude bands (i) 30°S-65°S, the entire Southern Ocean, (ii) 30°S-45°S, the sub-tropical Southern Ocean and (iii) 45°S-65°S, the mid-latitude Southern Ocean. (iii) Annual means of each variable (models and observations) plotted as zonal means, over 30°S-65°S, (iv) Scatter plots of seasonal mean downward (surface) and outgoing (TOA) longwave and short wave radiation as a function of~~,~~ total cloud cover, cloud liquid water path or cloud ice water path, calculated for the ~~3~~three regions outlined above. The data are regridded using a cubic interpolation method with the observations grid as target. Figure 15 provides an example diagnostic, with the top panel showing covariability of seasonal mean surface downward short wave radiation as a function of total cloud cover. To construct the figure, grid point values of cloud cover, for each season covering 30°S to 65°S, are saved into bins of 5% increasing cloud cover. For each grid point the corresponding seasonal mean radiation value is used to obtain a mean radiation flux for each cloud cover bin. The lower panel plots the fractional occurrence of seasonal mean cloud cover from CloudSat and model data for the same spatial and temporal averaging as used in the upper panel. Observations from CERES-EBAF radiation plotted against CloudSat cloud cover are compared to an example CMIP5 model. From the covariability plot we can diagnose whether models exhibit a similar dependency between incoming surface short wave radiation and cloud cover as seen in observations. We can further assess if there is a systematic bias in surface solar radiation and whether this bias occurs at specific values of cloud cover. Similar covariability plots are available for surface incoming longwave radiation and for TOA long and short wave radiation, plotted respectively against cloud cover, cloud liquid water path and cloud ice water path. Combining these diagnostics provides a comprehensive evaluation of simulated relationships between surface and TOA radiation fluxes and cloud variables.

4.2.3. Simulated tropical ocean climatology

An accurate representation of the tropical climate is fundamental for ESMs. The majority of solar energy received by the Earth is in the tropics and the potential for thermal emission of absorbed energy back to space is also largest in the tropics due to the high column concentrations of water vapor at low latitudes (Pierrehumbert, 1995; Stephens and Greenwald, 1991). Coupled interactions between equatorial SSTs, surface wind stress, precipitation and upper-ocean mixing are central to many tropical biases in ESMs. This is the case both with respect to the mean state and for key modes of variability, influenced by, or interacting with, the mean state (e.g., El Nino Southern Oscillation (ENSO), (Choi et al. (2011))). Such biases are often reflected in a “double ITCZ” seen in the majority of CMIP3 and CMIP5 CCMs (Li and Xie, 2014; Oueslati and Bellon, 2015). The double ITCZ bias, present in many ESMs, occurs when models fail to simulate a single, year round, ITCZ rainfall maximum north of the equator. Instead, an unrealistic secondary maximum in models south of the equator is present for some or all of the year. Such biases are particularly prevalent in the tropical Pacific, but can also occur in the Atlantic (Oueslati and Bellon, 2015). This double ITCZ is often accompanied by an overextension of the East Pacific equatorial cold tongue into the Central Pacific, collocated with a positive bias in easterly near-surface wind speeds and a shallow bias in ocean mixed layer depth (Lin, 2007). Such biases can directly impact the ability of an ESM to accurately represent ENSO variability (An et al., 2010; Guilyardi, 2006) and its potential sensitivity to climate change (Chen et al., 2015), with negative consequences for a range of simulated features, such as regional tropical temperature and precipitation variability, monsoon dynamics and ocean and terrestrial carbon uptake (Iguchi, 2011; Jones et al., 2001).

To assess such tropical biases with the ESMValTool, we have implemented a namelist with diagnostics motivated by the work of Li and Xie (2014) [*namelist_TropicalVariability.xml*]. In particular, we reproduce their Fig. 5 for models and observations/reanalyses, calculating equatorial mean (5°N-5°S), longitudinal sections of annual mean precipitation (pr), skin temperature (ts), horizontal winds (ua and va) and 925 hPa divergence (derived from the sum of the partial derivatives of the wind components extracted at the 925 hPa pressure level (that is du/dx + dv/dy). Latitude cross sections of the model variables are plotted for the equatorial Pacific, Indian and Atlantic Oceans with observational constraints provided by the TRMM-3B43-v7 for precipitation, the HadISST for SSTs, and ERA-interim reanalysis for temperature and winds. Latitudinal sections of absolute and normalized annual mean SST and precipitation are also calculated, spatially averaged for the three ocean basins. Normalization follows the procedure outlined in Fig. 1 of Li and Xie (2014) whereby values at each latitude are normalized by the tropical mean (20°N-20°S) value of the corresponding parameter (e.g., annual mean precipitation at a given location is divided by the 20°N-20°S annual mean value). Finally, to assess how models capture observed relationships between SST and precipitation we calculate the co-variability of precipitation against SST for specific regions of the tropical Pacific. This analysis includes calculation of the Mean Square Error (MSE) between model SST/precipitation and observational equivalents. A similar regridding procedure as for the Southern Hemisphere diagnostics is applied here, based on a cubic interpolation method and using the observations as target grid. The namelist as included in ESMValTool (v1.0) runs on all CMIP5 models. Figure 16 provides one example of the tropical climate diagnostics, with latitude cross sections of absolute and tropical normalized SST and precipitation from three CMIP5 models (HadGEM2-ES (Collins et al., 2011), MPI-ESM-LR and IPSL-CM5A-MR (Dufresne et al., 2013)) plotted against HadISST and TRMM data.

4.2.4. Sea ice

Sea ice is a key component of the climate system through its effects on radiation and seawater density. A reduction in sea ice area results in increased absorption of shortwave radiation, which warms the sea ice region and contributes to further sea ice loss. This process is often referred to as the sea ice albedo climate feedback which is part of the Arctic amplification phenomena (Curry, 2007). CMIP5 models tend to underestimate the ~~sharp~~ decline in summer Arctic sea ice extent observed by satellites during the last decades (Stroeve et al., 2012) which may be related to models' underestimation of the sea ice albedo feedback process ~~(Boé et al., 2009).~~(Boé et al., 2009). Conversely in the Antarctic, observations show a small increase in March sea ice extent while the CMIP5 models simulate a small decrease (Flato et al., 2013; Stroeve et al., 2012). It is therefore important that model sea-ice processes are evaluated and improvements regularly assessed. Caveats have been noted with respect to the limitations of using only sea ice extent as a metric of model performance (Notz et al., 2013) as the sea ice concentration, volume, and drift, sea ice thickness and surface albedo, as well as sea ice processes such as melt pond formation or the summer sea ice melt are all important sea ice related quantities. In addition the atmospheric forcings (e.g., wind, clouds, and snow) and ocean forcings (e.g., salinity and ocean transport) impact on the sea ice state and evolution.

In ESMValTool (v1.0) the sea ice namelist includes diagnostics that cover sea ice extent and concentration [*namelist_SeaIce.xml*], but work is underway to include other variables and processes in future releases. An example diagnostic produced by the sea ice namelist is given in Figure 17, which shows the timeseries of September Arctic sea ice extent from the CMIP5 historical simulations compared to observations from the National Snow and Ice Data Center (NSIDC) produced by combining concentration estimates created with the NASA Team algorithm and the Bootstrap algorithm (Meier et al., 2013; Peng et al., 2013) and SSTs from the HadISST data set, similar to Figure 9.24 of (Flato et al. (2013)). Sea ice extent is calculated as the total area (km^2) of grid cells over the Arctic or Antarctic with sea-ice concentrations (sic) of at least 15%. The sea ice namelist can also calculate the seasonal cycle of sea ice extent and polar stereographic contour and polar contour difference plots of Arctic and Antarctic sea ice concentration. For the latter diagnostic, data is regridded to a common 1°×1° grid using the patch recovery technique.

4.3. Detection of systematic biases in the physical climate: land

4.3.1. Continental dry bias

The representation of land surface processes and fluxes in climate models critically affects the simulation of near-surface climate over land. In particular, energy partitioning at the surface strongly influences surface temperature and it has been suggested that temperature biases in ESMs can be in part related to biases in evapotranspiration. The most notable feature in a majority of CMIP3 and CMIP5 models is a tendency to overestimate evapotranspiration globally ~~(~~(Mueller and Seneviratne, 2014)~~.~~(Mueller and Seneviratne, 2014).

A diagnostic to analyse the representation of evapotranspiration in ESMs has been included in the ESMValTool [*namelist_~~Evapotransport~~Evapotranspiration.xml*]. For comparison with the LandFlux-EVAL product (Mueller et al., 2013), the modelled surface latent heat flux (hfls) is converted to evapotranspiration units using the latent heat of vaporization. The diagnostic then produces lat-lon maps of absolute evapotranspiration as well as bias maps (model minus reference product~~).~~, after regridding data to the coarsest grid using area-conservative interpolation). In Fig. 18, the global pattern of monthly mean evapotranspiration is evaluated against the LandFlux-EVAL product. The evapotranspiration diagnostic is complemented by the Standardized Precipitation Index (SPI) diagnostic [*namelist_SPI.xml*], which gives a measure of drought intensity from an atmospheric perspective and can help relating biases in evapotranspiration to atmospheric causes such as the accumulated precipitation amounts. For each month, precipitation (pr) is summed over the preceding months (options for 3, 6 or 12-monthly SPI). Then a two-parameter Gamma distribution of cumulative probability is fitted to the strictly positive month sums, such that the probability of a non-zero precipitation sum being below a certain value x corresponds to Gamma(x). The shape and scale parameters of the gamma distribution are estimated with a maximum likelihood approach. Accounting for periods of no precipitation, occurring at a frequency q, the total cumulative probability distribution of a precipitation sum below x, H (x), becomes H (x) = q + (1 - q)*Gamma(x). In the last step, a precipitation sum x is assigned to its corresponding SPI value by computing the quantile q_N(0,1) of the standard normal distribution at probability H (x). The SPI of a precipitation sum x, thus, corresponds to the quantile of the standard normal distribution which is assigned by preserving the probability of the original precipitation sum, H (x). Mean and annual cycle are not meaningful since the SPI accounts for seasonality and transforms the data to a zero average in each month. Therefore the diagnostic focuses on lat-lon maps of annual or seasonal trends in SPI (unitless) ~~making comparison between~~when comparing models ~~and observation~~with observations.

4.3.2. Runoff

Evaluation of precipitation is a challenge due to potentially large errors and uncertainty in observed precipitation data (Biemans et al., 2009; Legates and Willmott, 1990). An alternative or additional option to the direct evaluation of precipitation over land (such as, e.g., included in the global precipitation evaluation in Sect. 4.1.2) is the evaluation of river runoff that can in principle be measured with comparatively small errors for most rivers. Routine measurements are performed for many large rivers, generating a large global database (e.g~~. available at the Global Runoff Data Centre (GRDC, Dümenil Gates et al. (2000)~~., available at the Global Runoff Data Centre (GRDC, Dümenil Gates et al. (2000))). The length of available time series, however, varies between the rivers, with large data gaps especially in recent years for many rivers. The evaluation of runoff against river gauge data can provide a useful independent measure of the simulated hydrological cycle. If both river flow and precipitation are given with reasonable accuracy, it will also provide an observational constraint on model surface evaporation, provided that the considered averaging time periods are long enough so that changes in surface water storages are negligible (Hagemann et al., 2013), e.g., by considering climatological means of 20 years or more. For present climate conditions ESMs often exhibit a dry and warm near-surface bias during summer over mid-latitude continents (Hagemann et al., 2004). Continental dry biases in precipitation exist in the majority of CMIP5 models over South America, the Mid-west of US, the Mediterranean region, Central and Eastern Europe, West and South Asia (Fig. ~~9.4 of Flato et al. (2013)~~4 and Fig. 9.4 of Flato et al. (2013)). These precipitation biases often transfer into dry biases in runoff, but sometimes dry biases in runoff can be caused by a too large evapotranspiration (Hagemann et al., 2013). In order to relate biases in runoff to biases in precipitation and evapotranspiration, the catchment oriented evaluation in this section considers biases in all three variables. This means that the respective variables are considered as spatially averages over the drainage basins of large rivers.

Beside bias maps, a set of diagnostics to produce basin-scale comparisons of runoff (mrro), evapotranspiration (evspsbl) and precipitation (pr) have also been implemented in ESMValTool [*namelist_runoff_et.xml*]. This namelist calculates biases in climatological annual means of the three variables for 12 large-scale catchments areas on different continents and for different climates. For total runoff, catchment averaged model values are compared to climatological long-term averages of GRDC observations. Due to the incompleteness of these station data, a year-to-year correspondence of data cannot be achieved so only climatological data are considered, as in

Hagemann et al. (2013). Simulated precipitation is compared to catchment-averaged WATCH forcing data based on ERA-Interim (WFDEI) data (Weedon et al., 2014) for the period 1979-2010. Evapotranspiration observations are estimated using the difference of the catchment-averaged WFDEI precipitation minus the climatological GRDC river runoff. As an example, Fig. 19 shows biases in runoff coefficient (runoff/precipitation) against the relative precipitation bias for the historical simulation of one of the CMIP5 models (MPI-ESM~~-1.1~~-LR).

4.4. Detection of biogeochemical biases: carbon cycle

4.4.1. Terrestrial biogeochemistry

A realistic representation of the global carbon cycle is a fundamental requirement for ESMs. In the past, climate models were directly forced by atmospheric CO_2 concentrations, but since CMIP5, ESMs are routinely forced by anthropogenic CO_2 emissions, the atmospheric concentration being inferred from the difference between these emissions and the ESM simulated land and ocean carbon sinks. These sinks are affected by atmospheric CO_2 and climate change, inducing feedbacks between the climate system and the carbon cycle (Arora et al., 2013; Friedlingstein et al., 2006). Quantification of these feedbacks is critical to estimate the future of these carbon sinks and hence atmospheric CO_2 and climate change (Friedlingstein et al., 2014).

The diagnostics implemented in ESMValTool to evaluate simulated terrestrial biogeochemistry are based on the study of Anav et al. (2013) and span several time-scales: climatological means, intra-annual (seasonal cycle), interannual and long-term trends [*namelist_anav13jclim.xml*]. Further extending these routines, ~~carbon cycle~~the diagnostics presented in Sect. 4.1.1 are also applied here to calculate quantitative performance metrics~~ from Anav et al. (2013) are implemented in *namelist_perfmetrics_CMIP5*.~~. These metrics assess how both the land and ocean biogeochemical components of ESMs reproduce different aspects of the land and ocean carbon cycle, with an emphasis on variables controlling the exchange of carbon between the atmosphere and these two reservoirs. The analysis indicates some level of compensating errors within the models. Selecting, within the namelist, several specific diagnostics to be applied to more key variables controlling the land or ocean carbon cycle, can help reducing the risk of missing such compensating errors. Figure 20 shows a portrait diagram similar to Fig. ~~3 but for seasonal carbon cycle metrics based on the point-wise RMSE against suitable reference data sets (see below). For annual mean trend diagnostics, such as those shown in Fig. 21, a PDF Skill Score metric is additionally implemented which compares the mean state and the interannual variability of a given variable at each grid point~~

~~by comparing the common area under both PDFs. The overlap of both PDFs provides a measure for the model ranking, with a perfect score of 1 meaning a full overlap of both PDFs (Anav et al. (2013), Eq.5~~3 of Anav et al. (2013) but for seasonal carbon cycle metrics against suitable reference data sets (see below).

For land, diagnostics of the land carbon sink net biosphere productivity (nbp) are essential. Although direct observations are not available, nbp can be estimated from atmospheric CO_2 inversions (JMA and TRANSCOM) and on the global scale combined with observation-based estimates of the oceanic carbon sink (fgco2 from GCP (Le Quéré et al., 2014)). In addition to net carbon fluxes, diagnostics for gross primary productivity of land (gpp), leaf area index (lai), vegetation (cVeg) and soil carbon pools (cSoil) are also implemented in the ESMValTool to assess possible error compensation in ESMs. Observation-based gpp estimates are derived from Model Tree Ensemble (MTE) upscaling data (Jung et al., 2009) from the network of eddy-covariance flux towers (FLUXNET, Beer et al. (2010). The leaf area index data set used for evaluation (LAI3g) is derived from the Global Inventory Modeling and Mapping Studies group (GIMMS) AVHRR normalized difference vegetation index (NDVI-017b) data (Zhu et al., 2013). Finally, cSoil and cVeg are assessed as mean annual values over different large sub-domains using the Harmonised World soil Database (HWSD, (Nachtergaele et al. (2012))) and the Olson based vegetation carbon data set (Gibbs, 2006; Olson et al., 1985).

4.4.2. Marine biogeochemistry

Marine biogeochemistry models form a core component of ESMs and require evaluation for multiple passive tracers. The increasing availability of quality-controlled global biogeochemical data sets for the historical period (e.g. Surface Ocean CO_2 Atlas Version 2 (SOCAT v2, Bakker et al. (2014)) provides further opportunity to evaluate model performance on multi-decadal timescales. Recent analyses of CMIP5 ESMs indicate that persistent biases exist in simulated biogeochemical variables, for instance as identified in ocean oxygen (Andrews et al., 2013) and carbon cycle (Anav et al., 2013) fields derived from CMIP5 historical experiments. Some systematic biases in biogeochemical tracers can be attributed to physical deficiencies within ocean models (see Section 4.~~3~~2), motivating further understanding of coupled physical-biogeochemical processes in the current generation of ESMs. For example, erroneous over oxygenation of subsurface waters within the MPI-ESM-LR CMIP5 model has been attributed to excess ventilation and vertical mixing in mid- to high-latitude regions (Ilyina et al., 2013).

A namelists is provided that includes diagnostics to support the evaluation of ocean biogeochemical cycles at global scales, as simulated by both ocean-only and coupled climate-carbon cycle ESMs [*namelist_GlobalOcean.xml*]. Supported input variables include surface partial pressure of CO_2 (spco2), surface chlorophyll concentration (chl), surface total alkalinity (talk) and dissolved oxygen concentration (o2). These variables provide an integrated view of model skill with regard to reproducing bulk marine ecosystem and carbon cycle properties. Observation-based reference data sets include SOCAT v2 and ETH-SOM-FFN (Landschützer et al., 2014a, b) for surface pCO_2 ~~(intPP),~~, Sea-viewing Wide Field-of-view Sensor (SeaWiFS) satellite data for surface chlorophyll (McClain et al., 1998), climatological data for total alkalinity (Takahashi et al., 2014), and World Ocean Atlas 2005 climatological data (WOA05) with in situ corrections following Bianchi et al. (2012) for dissolved oxygen. Diagnostics calculate contour plots for climatological distributions, inter-annual or inter-seasonal (e.g. JJAS) variability together with the difference between each model and a chosen reference data set. Such differences are calculated after regridding the data to the coarsest grid using an area-conservative interpolation. Monthly, seasonal or annual frequency time-series plots can also be produced either globally averaged or for a selected latitude-longitude range. Optional extensions include the ability to mask model data with the same coverage as observations, calculate anomaly fields, and to overlay trend lines, and running or multi-model means. Pre-processing routines are also included to accommodate native curvilinear grids, common in ocean model discretisation (see Section 4.~~3~~2.1), along with providing the ability to extract depth levels from 3-D input fields. An example plot is presented in Fig. 22, showing inter-annual variability in surface ocean pCO_2 as simulated by a subset of CMIP5 ESMs (BNU-ESM, HadGEM2-ES, GFDL-ESM2M), expressed as the standard deviation of de-trended annual averages for the period ~~1998 – 2011. The Representative Concentration Pathways (RCP) 4.5 CMIP5 model experiments are used to extend historical integrations beyond~~1992 – 2005~~ to facilitate comparison with~~. As an observation-based reference pCO_2 field~~ (~~, ETH SOM-FFN~~)~~, (1998-2011) is used, which extrapolates SOCAT v2 data (Bakker et al., 2014) using a 2-step neural network method. As described in Landschützer et al. (2014a), ETH SOM-FFN partitions monthly SOCAT v2 pCO_2 observations into discrete biogeochemical provinces by establishing common relationships between independent input parameters using a Self Organising Map (SOM). Non-linear input-target relationships, as derived for each biogeochemical province using a Feed-Forward Network (FFN) method, are then used to extrapolate observed pCO_2.

A diagnostic for oceanic Net Primary Production (NPP) is also implemented in ESMValTool for climatological annual mean and seasonal cycle, as well as for inter-annual variability over the 1986-2005 period [*namelist_anav13jclim.xml*]. Observations are derived from the SeaWiFS satellite chlorophyll data, using the Vertically Generalized Production Model (VGPM, Behrenfeld and Falkowski (1997)).~~ Finally, similarly to land carbon, the net air-sea CO_2 flux from ESMs (Fig. 21 right panels) is evaluated in terms of mean and interannual variations and climatological annual means over different zonally averaged domains using atmospheric inversions of the air-sea CO_2 flux as reference data (Gurney et al., 2003) and GCP estimates for the global ocean (Le Quéré et al., 2014).~~

4.5. Detection of biogeochemical biases: aerosols and trace gas chemistry

4.5.1. Tropospheric aerosols

Tropospheric aerosols play a key role in the Earth system and have a strong influence on climate and air pollution. The global aerosol distribution is characterized by a large spatial and temporal variability which makes its representation in ESMs particularly challenging (Ghan and Schwartz, 2007). In addition, aerosol interactions with radiation (direct aerosol effect (Schulz et al., 2006)) and with clouds (indirect aerosol effects (Lohmann and Feichter, 2005)) need to be accounted for. Model-based estimates of anthropogenic aerosol effects are still affected by large uncertainties, mostly due to an incorrect representation of aerosol processes (Kinne et al., 2006)~~. Myhre et al. (2013)~~. Myhre et al. (2013) report a substantial spread in simulated aerosol direct effects among 16 global aerosol models and attribute it to diversities in aerosol burden, aerosol optical properties and aerosol optical depth (AOD). Diversities in black carbon (BC) burden up to a factor of three, related to model disagreements in simulating deposition processes were also found by Lee et al. (2013). Model meteorology can be a source of diversity since it impacts on atmospheric transport and aerosol lifetime. This in turn relates to the simulated essential climate variables such as winds, humidity and precipitation (see Section 4.1). Large biases also exist in simulated aerosol indirect effects (IPCC, 2013) and are often a result of systematic errors in both model aerosol and cloud fields (see Section 4.1.6).

To assess current biases in global aerosol models, the aerosol namelist of the ESMValTool comprises several diagnostics to compare simulated aerosol concentrations and optical depth at the surface against station data, motivated by the work of ~~Pringle et al. (2010), ~~(Pozzer et al. (2012); Pringle et al. (2010)), Pozzer et al. (2012), and Righi et al. (2013) [*namelist_aerosol_CMIP5.xml*].

Diagnostics include time series of monthly or yearly mean aerosol concentrations, scatter plots with the relevant statistical indicators, and contour maps directly comparing model results against observations. ~~Comparison~~The comparison is performed considering collocated model and observations in space and time. In the current version of ESMValTool, these diagnostics are supplied with observational data from a wide range of station networks, including Interagency Monitoring of Protected Visual Environments (IMPROVE) and CASTNET (North America), European Monitoring and Evaluation Programme (EMEP, Europe) and the recently-established Asian network (EANET). The AERONET data are also available for evaluating aerosol optical depth in continental regions and in a few remote marine locations. For evaluating aerosol optical depth, we also use satellite data, the primary advantage of which is almost-global coverage, particularly over the oceans. Satellite data is however affected by uncertainties related to the algorithm used to process radiances into relevant geophysical state variables. The tool currently implements data from the Multi-angle Imaging SpectroRadiometer (MISR, Stevens and Schwartz (2012)), MODIS and the ESACCI-AEROSOL product (Kinne et al., 2015) which is a combination of ERS2-ATSR2 and ENVISAT-AATSR data. To calculate model biases against satellite data, regridding is performed using a bilinear interpolation to the coarsest grid. Aerosol optical depth time series over the ocean for the period 1850-~~2015~~2010 are shown in Fig. 23 for the CMIP5 models in comparison to MODIS and ESACCI-AEROSOL. Finally, more specific aerosol diagnostics have been implemented to compare aerosol vertical profiles of mass and number concentrations and aerosol size distributions, based on the evaluation work by Lauer et al. (2005) and Aquila et al. (2011). These diagnostics, however, use model quantities that were not part of the CMIP5 data request and therefore will not be discussed here.

4.5.2. Tropospheric trace gas chemistry and stratospheric ozone

In the past, climate models were forced with prescribed tropospheric and stratospheric ozone concentration, but since CMIP5 some ESMs include interactive chemistry and are capable of representing prognostic ozone (Eyring et al., 2013; Flato et al., 2013). This allows models to simulate important chemistry-climate interactions and feedback processes. Examples include the increase in oxidation rates in a warmer climate which leads to decreases in methane and its lifetime (Voulgarakis et al., 2013) or the increase in tropical upwelling (associated with the Brewer Dobson circulation) in a warmer climate and corresponding reductions in tropical lower stratospheric ozone as a result of faster transport and less time for ozone production (Butchart et al., 2010; Eyring et al., 2010). It is thus becoming important to evaluate the simulated atmospheric composition in ESMs. A common high bias in the Northern Hemisphere and a low bias in the Southern Hemisphere has been identified in tropospheric column ozone simulated by chemistry-climate models participating in the Atmospheric Chemistry Climate Model Intercomparison Project (ACCMIP), which could partly be related to deficiencies in the ozone precursor emissions (Young et al., 2013). Analysis of CMIP5 models with respect to trends in total column ozone show that the multi-model mean of the models with interactive chemistry is in good agreement with observations, but that significant deviations exist for individual models (Eyring et al., 2013; Flato et al., 2013). Large variations in stratospheric ozone in models with interactive chemistry drive large variations in lower stratospheric temperature trends. The results show that both ozone recovery and the rate of GHG increase determine future Southern Hemisphere summer-time circulation changes and are important to consider in ESMs (Eyring et al., 2013).

The namelists implemented in the ESMValTool to evaluate atmospheric chemistry ~~and the impact of stratospheric ozone changes on Southern Hemispheric surface climate~~ can reproduce the analysis of tropospheric ozone and precursors of Righi et al. (2015) [*namelist_righi15gmd_tropo3.xml*, *namelist_righi15gmd_Emmons.xml*] and the ~~studies by Eyring et al. (2006) and Eyring et al. (2013)~~study by Eyring et al. (2013) [*namelist_~~eyring06jgr.xml, namelist_~~eyring13jgr.xml*]. The calculation of the RMSE, mean bias, and Taylor diagrams (see Section 4.1.1) has been extended to tropospheric column ozone (derived from tro3 fields), ozone profiles (tro3) at selected levels, and surface carbon monoxide (vmrco) (see Righi et al. (2015) for details). This enables a consistent calculation of relative performance for the climate parameters and ozone, which is particularly relevant given that biases in climate can impact on biases in chemistry and vice versa. In addition, diagnostics that evaluate tropospheric ozone and its precursors (nitrogen oxides (vmrnox), ethylene (vmrc2h4), ethane (vmrc2h6), propene (vmrc3h6), propane (vmrc3h8) and acetone (vmrch3coch3)) are compared to the observational data of (Emmons et al. (2000)). A diagnostic to compare tropospheric column ozone from the CMIP5 historical simulations to Aura MLS/OMI observations (Ziemke et al., 2011) is also included and shown as an example in Fig. ~~24. For the stratosphere, total column ozone (toz) and processes-oriented diagnostics from Eyring et al. (2006) are implemented that include the seasonal cycle of temperature at 100 hPa and the correlation of the heat flux at 100 hPa (vt100) versus temperatures (ta) at 50 hPa, both evaluated with ERA-40 data (Uppala et al., 2005), vertical and latitudinal profiles of inorganic chlorine (cly), methane (ch4), water vapour (h2o) and time-height sections of water vapour mixing ratio (i.e., the tape recorder) evaluated with satellite data from HALOE (Grooß and Russell Iii, 2005), and age of air (age)~~

~~evaluated against satellite measurements of HF and HCl from HALOE [Anderson et al., 2000] and other sources (see Eyring et al. (2006) for details). Figure 25 shows the CMIP5 total column ozone time series compared to five different observational data sets: ground-based measurements (updated from Fioletov et al. (2002)), NASA TOMS/OMI/SBUV(/2) merged satellite data (Stolarski and Frith, 2006),~~ 24. This diagnostic also remaps the data to the coarsest grid using local area averaging in order to calculate differences. For the stratosphere, total column ozone (toz) diagnostics are implemented. As an example, Figure 25 shows the CMIP5 total column ozone time series compared to the NIWA combined total column ozone database (Bodeker et al., 2005)~~, Solar Backscatter Ultraviolet (SBUV, SBUV/2) retrievals (updated from Miller et al. (2002)), and GOME/SCIA/GOME-2 (Loyola and Coldewey-Egbers, 2012).~~.

4.6. Linking model performance to projections

The relatively new research field of emergent constraints aims to link model performance evaluation with future projection feedbacks. An emergent constraint refers to the use of observations to constrain a simulated future Earth system feedback. It is referred to as emergent, because a relationship between a simulated future projection feedback and an observable element of climate variability emerges from an ensemble of ESM projections, potentially providing a constraint on the future feedback. Emergent constraints can help focus model development and evaluation onto processes underpinning uncertainty in the magnitude and spread of future Earth system change. Systematic model biases in certain forced modes, such as the seasonal cycle of snow cover or inter-annual variability of tropical land CO_2 uptake appear to project in an understandable way onto the spread of future climate change feedbacks resulting from these phenomena (Cox et al., 2013; Hall and Qu, 2006; Wenzel et al., 2014).

To reproduce the analysis of Wenzel et al. (2014) that provides an emergent constraint on future tropical land carbon uptake, a namelist is included into ESMValTool (v1.0) that performs emergent constraint analysis of the carbon cycle-climate feedback parameter (γ_{LT}) (Friedlingstein et al. 2006; Cox et al. 2013) [*namelist_wenzel14jgr.xml*]. This namelist only considers the CMIP5 ESMs that provided the necessary output for the analysis. This criterion precludes most CMIP5 models and only seven ESMs are included here. The namelist includes diagnostics which analyse the short-term sensitivity of atmospheric CO_2 to temperature variability on interannual time scales (γ_{IAV}) for models and observations, as well as diagnostics for γ_{LT} from the models. The observed sensitivity γ_{IAV} is calculated by summing land (nbp) and ocean (fgco2) carbon fluxes which are correlated to tropical near-surface air temperature (tas). Results from historical model simulations are compared to observational based estimates of carbon fluxes from the Global Carbon project (GCP, (Le Quéré et al., 2014)) and reanalysis temperature data from the NOAA National Climate Data Center (NCDC, ~~Smith et al. (2008)~~). For diagnosing γ_{LT} from the models, nbp from idealized fully coupled and biochemically coupled simulations are used as well as tas from fully coupled idealized simulations (see Fig. 26). Emergent constraints of this type help to understand some of the underlying processes controlling future projection sensitivity and offer a promising approach to reduce uncertainty in multi-model climate projections.

5. Use of the ESMValTool in the model development cycle and evaluation workflow

5.1. Model development

As new model versions are developed, standardized diagnostics suites as presented here allow model developers to compare their results against previous versions of the same model or against other, e.g. CMIP, models. Such analyses help to identify different aspects in a model that have either improved or degraded as a result of a particular model development. The benchmarking of ESMs using performance metrics (see Section 4.1.1) provides an overall picture of the quality of the simulation, whereas process-oriented diagnostics help determine whether the simulation quality improvements are for the correct underlying physical reasons and point to paths for further model improvement.

The ESMValTool is intended to support modelling centres with quality control of their CMIP DECK experiments and the CMIP6 historical simulation, as well as other experiments ~~related to the individual Model Intercomparison Projects (MIPs) that are part of CMIP6.~~from CMIP6-Endorsed Model Intercomparison Projects (Eyring et al., 2015). A significant amount of institutional resources go into running, post-processing, and publishing model results from such experiments. It is important that centres can easily identify and correct potential errors in this process. The standardized analyses contained in the ESMValTool can be used to monitor the progress of CMIP experiments. While the tool is designed to accommodate a wide range of time axes and configurations, and many of the diagnostics may be run on control or future climate experiments, ESMValTool (v1.0) is largely targeted to evaluate AMIP and the CMIP historical simulations.

5.2. Integration into modelling workflows

The ESMValTool can be run as a stand-alone tool, or ~~be~~ integrated into existing modelling workflows. The primary challenge is to provide CF/CMOR compliant data. Not all modelling centres produce CF/CMOR compliant data directly as part of their workflow although we note that more are doing so as the potential benefits are being realized. For many groups conversion to CF/CMOR standards involves significant post-processing of native model output. This may require some groups to perform analysis via the ESMValTool on their model output after conversion to CF/CMOR, or to create intermediate "CMOR-like" versions of the data. Users who wish to use native model output can take advantage of the reformatting routine flexibility (see Section ~~2.~~3) to create scripts that convert this data into the CF/CMOR standard. As an example, reformat scripts for the NOAA-GFDL models and the EMAC model are included with the initial release. These scripts are used to convert the native model output for direct use with the ESMValTool. The reformatting routine capability may provide an alternative to more expensive and complete "CMORization" processes that are usually required to formally publish model data on the ESGF.

5.3. Running the ESMValTool alongside the ESGF

Large international model inter-comparison projects ~~(~~such as CMIP~~)~~ stimulated the development of a globally distributed federation of data providers, supporting common data provisioning policies and infrastructures. ESGF is an international open source effort to establish a distributed data and computing platform, enabling world wide access to Peta- (in the future Exa-) byte scale scientific climate data. Data can be searched via a globally distributed search index with access possible via HTTP, OpenDAP and GridFTP. To efficiently run the ESMValTool on CMIP model data and observations alongside the ESGF, the necessary data hosted by the ESGF has to be made locally accessible at the site where ESMValTool is executed. ~~There are two possibilities (which can be exploited in parallel) to accomplish this. The first is to configure ESMValTool to use locally available data which is independently managed in a local ESGF data pool (replica and published files). The second option is to download files remotely from the ESGF and cache them on the user's local system, under the control of a 'Data Manager', which may be part of the ESMValTool software, or existing third party software under user (or local administrator) control. Larger ESGF sites often act as replica centres maintaining a large ESGF replica pool (e.g., DKRZ, BADC, IPSL, PCMDI) and thus can effectively exploit the first option. Others can rely on the ESMValTool Data Manager to download and maintain a download cache of required input data sets. Both options~~

~~require configuration of ESMValTool to use data organized in a hierarchical directory tree, organized following the CMIP conventions. Figure 27 provides a schematic overview of the coupling of the ESMValTool to the ESGF. As mentioned, ESMValTool uses a standard namelist written in XML to define models and variables to be analysed. If the ESGF enabled ESMValTool software is running on an ESGF node, the Data Manager can use information held in the namelist to locate the correct file within the node's local data pool. If the file is unavailable there, or ESMValTool is not running on an ESGF node, the Data Manager can instead use namelist information to locate the file in the local download cache (see above). If files are not available they will be downloaded and stored in the download cache. Using a cache avoids downloading of files more than once. Thus using the Data Manager, which is currently being developed, the ESMValTool is decoupled from the distributed ESGF data infrastructure, which acts as the data source for local copies of the required files~~There are various ways this might be achieved. One possibility is to run ESMValTool separately at each site holding datasets required by the analysis, then combine the results. However, this is limited by the extent to which calculations can be performed without requiring data from another site. A more practical possibility is running ESMValTool alongside a large store of replica datasets gathered from across the ESGF, so that all the required data are in one location. Certain large ESGF sites (e.g., DKRZ, BADC, IPSL, PCMDI) provide replica dataset stores, and ESMValTool has been run in such a way at several of these sites.

Replica dataset stores do not provide a complete solution however, as it is impossible to replicate all ESGF datasets at one site, so circumstances will arise when one or more required datasets are not available locally. The obvious solution is to download these datasets from elsewhere in the ESGF, and store them locally whilst the analysis is carried out. The indexed search facility provided by the ESGF makes it easy to identify the download URL of such 'remote' datasets, and a prototype of ESMValTool (not included in v1.0) has been developed that performs this search automatically using esgf-pyclient[1]. If the search is successful, the prototype provides the user with the URL of each file in the dataset, and the user (or system administrator) is then responsible for performing the download. The workflow of this prototype is illustrated in Figure 27. It is possible that the fully automated downloading of remote ESGF datasets may be provided by a future version of ESMValTool, but for now it is preferable for a human to manage the process due to large size of the

[1] https://pypi.python.org/pypi/esgf-pyclient

Form

files involved. A more complete coupling to the ESGF was originally planned for version 1.0 but was not possible due to the long down period of the ESGF.

6. Summary and Outlook

The Earth System Model ~~eValuation~~Evaluation Tool (ESMValTool) is a diagnostics package for routine evaluation of Earth System Models (ESMs) ~~against~~with observations and reanalyses data or for comparison with results from other models. The ESMValTool has been developed to facilitate the evaluation of complex ESMs at individual modelling centres and to help streamline model evaluation standards within CMIP. Priorities to date that are included in ESMValTool (v1.0) described in this paper~~,~~ concentrate on selected systematic biases that were a focus of the European Commission's 7th Framework Programme "Earth system Model Bias Reduction and assessing Abrupt Climate change (EMBRACE) project, the DLR Earth System Model Evaluation (ESMVal) project and other collaborative projects, in particular: performance metrics for selected ECVs, coupled tropical climate variability, monsoons, Southern Ocean processes, continental dry biases and soil hydrology-climate interactions, atmospheric CO_2 budgets, ozone, and tropospheric aerosol. We have applied the bulk of the diagnostics of ESMValTool (v1.0) to the entire set of CMIP5 historical or AMIP simulations. The namelist on emergent constraints for the carbon cycle has been additionally applied to idealized carbon cycle experiments and the emission driven RCP 8.5 simulations.

ESMValTool (v1.0) can be used to compare new model simulations against CMIP5 models and observations for the selected scientific themes much faster than this was possible before. Model groups, who wish to do this comparison before submitting their CMIP6 ~~Historical Simulation~~historical simulations or AMIP ~~experiment~~experiments to the ESGF can do so since the tool is provided as open source software. In order to run the tool locally, observations need to be downloaded and for tiers 2 and 3 reformatted with the help of the reformatting scripts that are included. Model output needs to be either in CF compliant NetCDF or a reformatting routine needs to be written by the modelling group, following given examples for EMAC, GFDL models, and NEMO.

Users of the ESMValTool (v1.0) results need to be aware that ESMValTool (v1.0) only includes a subset of the wide behaviour of model performance that the community aims to characterize. The results of running the ESMValTool need to be interpreted accordingly. Over time, the ESMValTool will be extended with additional diagnostics and performance metrics. A particular focus will be to integrate additional diagnostics that can reproduce the analysis of the climate model evaluation chapter of IPCC AR5 (Flato et al., 2013) as well as the projection chapter (Collins et al., 2013). We will also extend the tool with diagnostics to quantify forcings and feedbacks in the CMIP6 simulations and to calculate metrics such as the equilibrium climate sensitivity (ECS), transient climate response (TCR), and the transient climate response to cumulative carbon emissions (TCRE) ~~from the idealized CMIP experiments (IPCC, 2013). While inclusion of these diagnostics is straight forward~~(IPCC, 2013). While inclusion of these diagnostics is straightforward, the evaluation of processes and phenomena to improve understanding about the sources of errors and uncertainties in models that we also plan to enhance remains a scientific challenge. The field of emergent constraints remains in its infancy and more research is required how to better link model performance to projections (Flato et al., 2013). In addition, an improved consideration of the interdependency in the evaluation of a multi-model ensemble (Sanderson et al., 2015a, b) as well as internal variability in ESM evaluation is required.

A critical aspect in ESM evaluation is the availability of consistent, error-characterized global and regional Earth observations, as well as accurate globally gridded reanalyses that are constrained by assimilated observations. Additional or longer records of observations and reanalyses will be used as they become available, with a focus on using obs4MIPs - including new contributions from the European Space Agency's Climate Change Initiative (ESA CCI) - and ana4MIPs data. The ESMValTool can consider observational uncertainty in different ways, e.g. through the use of more than one observational data set to directly evaluate the models, by showing the difference between the reference data set and the alternative observations, or by including an observed uncertainty ensemble that spans the observed uncertainty range (e.g~~.~~., available for the surface temperature data set compiled for HadISST). Often the uncertainties in the observations are not readily available. Reliable and robust error characterization/estimation of observations is a high priority throughout the community, and obs4MIPs and other efforts that create data sets for model evaluation should encourage the inclusion of such uncertainty estimates as part of each data set.

The ESMValTool will be contributed to the analysis code catalogue being developed by the WGNE/WGCM climate model metrics panel~~ (http://www-metrics-panel.llnl.gov/wiki).~~. The purpose of this catalogue is to make the diversity of existing community-based analysis capabilities more accessible and transparent, and ultimately for developing solutions to ensure they can be readily applied to the CMIP DECK and the CMIP6 historical simulation in a coordinated way. We are currently exploring options to interface with complimentary efforts, e.g. ~~the PCMDI metrics package (Gleckler et al., EOS, 2015)~~the PCMDI Metrics Package (PMP, Gleckler et al. (2016)) and the Auto-Assess package that is under development at the UK Met Office. An international strategy for organising and presenting CMIP results produced by various diagnostic tools is needed, and this will be a priority for the WGNE/WGCM climate metrics panel in collaboration with the CMIP Panel (http://www.wcrp-climate.org/index.php/wgcm-cmip/about-cmip).

This paper presents ESMValTool (v1.0) which allows users to repeat all the analyses shown. Additional updates and improvements will be included in subsequent versions of the software, which are planned to be released on a regular basis. The ESMValTool works on CMIP5 simulations and, given CMIP DECK and CMIP6 simulations will be in a similar format, it will be straightforward to run the package on these simulations. A limiting factor at present is the need to download all data to a local cache. This limitation has spurred the development allowing ESMValTool to run alongside the ESGF at one of the data nodes. An initial attempt to couple the tool to the ESGF has been made, but ~~further improvements are required.~~this is still at prototype stage (see Section 5.3). An additional limiting factor is that the model output from all CMIP models has to be mirrored to the ESGF data node where the tool is installed. This is facilitated by providing a listing of the variables and time frequencies that are used in ESMValTool (v1.0) which uses a significantly smaller volume than the data request for the CMIP DECK and CMIP6 simulations will include. This reduced set of data could be mirrored with priority.

Several technical improvements are required to make the software package more efficient. One current limitation is the lack of a parallelization. Given the huge amount of data involved in a typical CMIP analysis, this can be highly CPU-time-intensive when performed on a single processor. In future releases, the possibility of parallelizing the tool will be explored. Additional development work is ongoing to create a more flexible pre-processing framework, which will include operations like ensemble-averaging and regridding to the current reformatting procedures as well as an improved coupling to the ESGF. Here, future versions of the ESMValTool will build as much as possible on existing efforts for the backend that reads and reformats data. In this regard it would be helpful if an application programming interface (API) could be defined for example by the WGCM Infrastructure Panel (WIP) that allows for flexible integration of diagnostics across different tools and programming languages in CMIP to this backend.

We aim to move ESM evaluation beyond the state-of-the-art by investing in operational evaluation of physical and biogeochemical aspects of ESMs, process-oriented evaluation and by identifying processes most important to the magnitude and uncertainty of future projections. Our goal is to support ~~CMIP DECK and CMIP6~~model evaluation in CMIP6 by contributing the ESMValTool as one of the standard documentation functions and by running it alongside the ESGF. In collaboration with similar efforts, we aim for a routine evaluation that provides a comprehensive documentation of broad aspects of model performance and its evolution over time and to make evaluation results available at a timescale that was not possible in CMIP5. This routine evaluation is not meant to replace further in-depth analysis of model performance and can to date not strongly reduce uncertainties in global climate sensitivity which remains an active area of research. However, the ability to routinely perform such evaluation will drive the quality and realism of ESMs forward and will leave more time to develop innovative process-oriented diagnostics - especially those related to feedbacks in the climate system that link to the credibility of model projections.

7. Code availability

ESMValTool ~~VERSION 1.0~~ (v1.0) ~~that is described in this paper will be made available from the ESMValTool web-page at http://www.pa.op.dlr.de/ESMValTool via a tar-file with a Digital Object Identifier (doi) assigned. ESMValTool (v1.0) will be~~is released under the Apache License, VERSION 2.0~~ and citation~~. The latest version of ~~this paper is~~the ESMValTool is available from the ESMValTool webpage at http://www.esmvaltool.org/. Users who apply the Software resulting in presentations or papers are kindly ~~requested upon use~~asked to cite this paper alongside with the Software doi (doi:10.17874/ac8548f0315) and version number. In addition, ESMValTool will be further developed in a version controlled repository that is accessible only to the development team. Regular releases are planned for the future. The wider climate community is encouraged to contribute to this effort and to join the ESMValTool development team for contribution of additional more in-depth diagnostics for ESM evaluation. A wiki page for the development that describes ongoing developments is also available. Interested users and developers are welcome to contact the lead author.

Acknowledgements

The development of ~~the~~ ESMValTool (v1.0) was funded by the European Commission's 7th Framework Programme, under Grant Agreement number 282672, the "Earth system Model Bias Reduction and assessing Abrupt Climate change (EMBRACE)" project and the DLR "Earth System

Model Validation (ESMVal)” ~~project.~~and “Klimarelevanz von atmosphärischen Spurengasen, Aerosolen und Wolken: Auf dem Weg zu EarthCARE und MERLIN (KliSAW)” projects. In addition, financial support for the development of ~~the~~ ESMValTool (v1.0) was provided by ESA’s Climate Change Initiative Climate Modelling User Group (CMUG). We acknowledge the World Climate Research Program’s (WCRP’s) Working Group on Coupled Modelling (WGCM), which is responsible for CMIP, and we thank the climate modelling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. ~~We thank C.~~We thank Björn Brötz (DLR, Germany) for his help with the release of the ESMValTool and Clare Enright (UEA, UK) for support with development of the ocean biogeochemistry diagnostics. We are grateful to Patrick Jöckel (DLR, Germany~~) and~~), Ron Stouffer (GFDL, USA) and to the two anonymous referees for their constructive comments on the manuscript.

References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J., Arkin, P., and Nelkin, E.: The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979–Present), J Hydrometeorol, 4, 1147-1167, 2003.

Alaka, G. J. and Maloney, E. D.: The Influence of the MJO on Upstream Precursors to African Easterly Waves, J Climate, 25, 3219-3236, 2012.

An, S. I., Ham, Y. G., Kug, J. S., Timmermann, A., Choi, J., and Kang, I. S.: The Inverse Effect of Annual-Mean State and Annual-Cycle Changes on ENSO, J Climate, 23, 1095-1110, 2010.

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung, M., Myneni, R., and Zhu, Z.: Evaluating the Land and Ocean Components of the Global Carbon Cycle in the CMIP5 Earth System Models, J Climate, 26, 6801-6843, 2013.

Andrews, O. D., Bindoff, N. L., Halloran, P. R., Ilyina, T., and Le Quere, C.: Detecting an external influence on recent changes in oceanic oxygen using an optimal fingerprinting method, Biogeosciences, 10, 1799-1813, 2013.

Annamalai, H., Hamilton, K., and Sperber, K. R.: The South Asian summer monsoon and its relationship with ENSO in the IPCC AR4 simulations, J Climate, 20, 1071-1092, 2007.

Antonov, J. I., Seidov, D., Boyer, T. P., Locarnini, R. A., Mishonov, A. V., Garcia, H. E., Baranova, O. K., Zweng, M. M., and Johnson, D. R.: World Ocean Atlas 2009, Volume 2: Salinity. In: NOAA Atlas NESDIS 69, Levitus, S. (Ed.), U.S. Government Printing Office, Washington, D.C., 2010.

Aquila, V., Hendricks, J., Lauer, A., Riemer, N., Vogel, H., Baumgardner, D., Minikin, A., Petzold, A., Schwarz, J. P., Spackman, J. R., Weinzierl, B., Righi, M., and Dall'Amico, M.: MADE-in: a new aerosol microphysics submodel for global simulation of insoluble particles and their mixing state, Geosci Model Dev, 4, 325-355, 2011.

Arora, V. K., Boer, G. J., Friedlingstein, P., Eby, M., Jones, C. D., Christian, J. R., Bonan, G., Bopp, L., Brovkin, V., Cadule, P., Hajima, T., Ilyina, T., Lindsay, K., Tjiputra, J. F., and Wu, T.: Carbon-Concentration and Carbon-Climate Feedbacks in CMIP5 Earth System Models, J Climate, 26, 5289-5314, 2013.

Ashok, K., Guan, Z. Y., Saji, N. H., and Yamagata, T.: Individual and combined influences of ENSO and the Indian Ocean Dipole on the Indian summer monsoon, J Climate, 17, 3141-3155, 2004.

Aumann, H. H., Chahine, M. T., Gautier, C., Goldberg, M. D., Kalnay, E., McMillin, L. M., Revercomb, H., Rosenkranz, P. W., Smith, W. L., Staelin, D. H., Strow, L. L., and Susskind, J.: AIRS/AMSU/HSB on the Aqua mission: design, science objectives, data products and processing system, EEE Trans. Geosci. and Remote Sensing, 41, 253-264, 2003.

Bakker, D. C. E., Pfeil, B., Smith, K., Hankin, S., Olsen, A., Alin, S. R., Cosca, C., Harasawa, S., Kozyr, A., Nojiri, Y., O'Brien, K. M., Schuster, U., Telszewski, M., Tilbrook, B., Wada, C., Akl, J., Barbero, L., Bates, N. R., Boutin, J., Bozec, Y., Cai, W. J., Castle, R. D., Chavez, F. P., Chen, L., Chierici, M., Currie, K., de Baar, H. J. W., Evans, W., Feely, R. A., Fransson, A., Gao, Z., Hales, B., Hardman-Mountford, N. J., Hoppema, M., Huang, W. J., Hunt, C. W., Huss, B., Ichikawa, T., Johannessen, T., Jones, E. M., Jones, S. D., Jutterström, S., Kitidis, V., Körtzinger, A., Landschützer, P., Lauvset, S. K., Lefèvre, N., Manke, A. B., Mathis, J. T., Merlivat, L., Metzl, N., Murata, A., Newberger, T., Omar, A. M., Ono, T., Park, G. H., Paterson, K., Pierrot, D., Ríos, A. F., Sabine, C. L., Saito, S., Salisbury, J., Sarma, V. V. S. S., Schlitzer, R., Sieger, R., Skjelvan, I., Steinhoff, T., Sullivan, K. F., Sun, H., Sutton, A. J., Suzuki, T., Sweeney, C., Takahashi, T., Tjiputra, J., Tsurushima, N., van Heuven, S. M. A. C., Vandemark, D., Vlahos, P., Wallace, D. W. R., Wanninkhof, R., and Watson, A. J.: An update to the Surface Ocean CO2 Atlas (SOCAT version 2), Earth Syst. Sci. Data, 6, 69-90, 2014.

Barkstrom, B. R.: The Earth Radiation Budget Experiment (ERBE), B Am Meteorol Soc, 65, 1170-1185, 1984.

Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Schamm, K., Schneider, U., and Ziese, M.: A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present, Earth Syst. Sci. Data, 5, 71-99, 2013.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rodenbeck, C., Arain, M. A., Baldocchi, D., Bonan, G. B., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K. W., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, F. I., and Papale, D.: Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate, Science, 329, 834-838, 2010.

Behrenfeld, M. J. and Falkowski, P. G.: Photosynthetic rates derived from satellite-based chlorophyll concentration, Limnol Oceanogr, 42, 1-20, 1997.

Bianchi, D., Dunne, J. P., Sarmiento, J. L., and Galbraith, E. D.: Data-based estimates of suboxia, denitrification, and N2O production in the ocean and their sensitivities to dissolved O-2, Global Biogeochem Cy, 26, 2012.

Biasutti, M.: Forced Sahel rainfall trends in the CMIP5 archive, Journal of Geophysical Research: Atmospheres, 118, 1613-1623, 2013.

Biemans, H., Hutjes, R. W. A., Kabat, P., Strengers, B. J., Gerten, D., and Rost, S.: Effects of Precipitation Uncertainty on Discharge Calculations for Main River Basins, J Hydrometeorol, 10, 1011-1025, 2009.

Bodas-Salcedo, A., Williams, K. D., Ringer, M. A., Beau, I., Cole, J. N. S., Dufresne, J. L., Koshiro, T., Stevens, B., Wang, Z., and Yokohata, T.: Origins of the Solar Radiation Biases over the Southern Ocean in CFMIP2 Models, J Climate, 27, 41-56, 2014.
Bodeker, G. E., Shiona, H., and Eskes, H.: Indicators of Antarctic ozone depletion, Atmos Chem Phys, 5, 2603-2615, 2005.
Boé, J., Hall, A., and Qu, X.: Current GCMs' Unrealistic Negative Feedback in the Arctic, J Climate, 22, 4682-4695, 2009.
Bollasina, M. and Nigam, S.: Indian Ocean SST, evaporation, and precipitation during the South Asian summer monsoon in IPCC-AR4 coupled simulations, Clim Dynam, 33, 1017-1032, 2009.
Bollasina, M. A. and Ming, Y.: The general circulation model precipitation bias over the southwestern equatorial Indian Ocean and its implications for simulating the South Asian monsoon, Clim Dynam, 40, 823-838, 2013.
Boucher, O., D. Randall, P. Artaxo, C. Bretherton, G. Feingold, P. Forster, V.-M. Kerminen, Y. Kondo, H. Liao, U. Lohmann, P. Rasch, S.K. Satheesh, S. Sherwood, Stevens, B., and Zhang, X. Y.: Clouds and Aerosols. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Stocker, T. F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (Ed.), Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.
Butchart, N., Cionni, I., Eyring, V., Shepherd, T. G., Waugh, D. W., Akiyoshi, H., Austin, J., Brühl, C., Chipperfield, M. P., Cordero, E., Dameris, M., Deckert, R., Dhomse, S., Frith, S. M., Garcia, R. R., Gettelman, A., Giorgetta, M. A., Kinnison, D. E., Li, F., Mancini, E., McLandress, C., Pawson, S., Pitari, G., Plummer, D. A., Rozanov, E., Sassi, F., Scinocca, J. F., Shibata, K., Steil, B., and Tian, W.: Chemistry–Climate Model Simulations of Twenty-First Century Stratospheric Climate and Circulation Changes, J Climate, 23, 5349-5374, 2010.
Chen, L., Li, T., and Yu, Y. Q.: Causes of Strengthening and Weakening of ENSO Amplitude under Global Warming in Four CMIP5 Models, J Climate, 28, 3250-3274, 2015.
Chen, W. T., Woods, C. P., Li, J. L. F., Waliser, D. E., Chern, J. D., Tao, W. K., Jiang, J. H., and Tompkins, A. M.: Partitioning CloudSat ice water content for comparison with upper tropospheric ice in global atmospheric models, J Geophys Res-Atmos, 116, 2011.
Cherchi, A. and Navarra, A.: Influence of ENSO and of the Indian Ocean Dipole on the Indian summer monsoon variability, Clim Dynam, 41, 81-103, 2013.
Cheruy, F., Dufresne, J. L., Hourdin, F., and Ducharne, A.: Role of clouds and land-atmosphere coupling in midlatitude continental summer warm biases and climate change amplification in CMIP5 simulations, Geophys Res Lett, 41, 6493-6500, 2014.
Choi, J., An, S. I., Kug, J. S., and Yeh, S. W.: The role of mean state on changes in El Nio's flavor, Clim Dynam, 37, 1205-1215, 2011.
Collins, M., R. Knutti, J. Arblaster, J.-L. Dufresne, T. Fichefet, P. Friedlingstein, X. Gao, W.J. Gutowski, T. Johns, G. Krinner, M. Shongwe, C. Tebaldi, A.J. Weaver, and Wehner, M.: Long-term Climate Change: Projections, Commitments and Irreversibility. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Stocker, T. F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (Ed.), Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.
Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., Hughes, J., Jones, C. D., Joshi, M., Liddicoat, S., Martin, G., O'Connor, F., Rae, J., Senior, C., Sitch, S., Totterdell, I., Wiltshire, A., and Woodward, S.: Development and evaluation of an Earth-System model-HadGEM2, Geosci Model Dev, 4, 1051-1075, 2011.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Bronnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, O., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D., and Worley, S. J.: The Twentieth Century Reanalysis Project, Q J Roy Meteor Soc, 137, 1-28, 2011.

Connolley, W. M. and Bracegirdle, T. J.: An Antarctic assessment of IPCC AR4 coupled models, Geophys. Res. Lett., 34, L22505, 2007.

Cook, K. H. and Vizy, E. K.: Coupled model simulations of the west African monsoon system: Twentieth- and Twenty-First-century simulations, J Climate, 19, 3681-3703, 2006.

Cox, P. M., Pearson, D., Booth, B. B., Friedlingstein, P., Huntingford, C., Jones, C. D., and Luke, C. M.: Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability, Nature, 494, 341-344, 2013.

Curry, C. L.: Modeling the soil consumption of atmospheric methane at the global scale, Global Biogeochem. Cycles, 21, GB4012, 2007.

Danabasoglu, G., Bates, S. C., Briegleb, B. P., Jayne, S. R., Jochum, M., Large, W. G., Peacock, S., and Yeager, S. G.: The CCSM4 Ocean Component, J Climate, 25, 1361-1389, 2012.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Holm, E. V., Isaksen, L., Kallberg, P., Kohler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q J Roy Meteor Soc, 137, 553-597, 2011.

Deser, C., Alexander, M. A., Xie, S. P., and Phillips, A. S.: Sea Surface Temperature Variability: Patterns and Mechanisms, Annu Rev Mar Sci, 2, 115-143, 2010.

Deser, C., Knutti, R., Solomon, S., and Phillips, A. S.: Communication of the role of natural variability in future North American climate, Nat Clim Change, 2, 775-779, 2012.

Deser, C., Phillips, A. S., Alexander, M. A., and Smoliak, B. V.: Projecting North American Climate over the Next 50 Years: Uncertainty due to Internal Variability*, J Climate, 27, 2271-2296, 2014.

Dong, S., Sprintall, J., Gille, S. T., and Talley, L.: Southern Ocean mixed-layer depth from Argo float profiles, J. Geophys. Res., 113, C06013, 2008.

Donner, L. J., Wyman, B. L., Hemler, R. S., Horowitz, L. W., Ming, Y., Zhao, M., Golaz, J. C., Ginoux, P., Lin, S. J., Schwarzkopf, M. D., Austin, J., Alaka, G., Cooke, W. F., Delworth, T. L., Freidenreich, S. M., Gordon, C. T., Griffies, S. M., Held, I. M., Hurlin, W. J., Klein, S. A., Knutson, T. R., Langenhorst, A. R., Lee, H. C., Lin, Y. L., Magi, B. I., Malyshev, S. L., Milly, P. C. D., Naik, V., Nath, M. J., Pincus, R., Ploshay, J. J., Ramaswamy, V., Seman, C. J., Shevliakova, E., Sirutis, J. J., Stern, W. F., Stouffer, R. J., Wilson, R. J., Winton, M., Wittenberg, A. T., and Zeng, F. R.: The Dynamical Core, Physical Parameterizations, and Basic Simulation Characteristics of the Atmospheric Component AM3 of the GFDL Global Coupled Model CM3, J Climate, 24, 3484-3519, 2011.

Dufresne, J. L., Foujols, M. A., Denvil, S., Caubel, A., Marti, O., Aumont, O., Balkanski, Y., Bekki, S., Bellenger, H., Benshila, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., Noblet, N., Duvel, J. P., Ethé, C., Fairhead, L., Fichefet, T., Flavoni, S., Friedlingstein, P., Grandpeix, J. Y., Guez, L., Guilyardi, E., Hauglustaine, D., Hourdin, F., Idelkadi, A., Ghattas, J., Joussaume, S., Kageyama, M., Krinner, G., Labetoulle, S., Lahellec, A., Lefebvre, M. P., Lefevre, F., Levy, C., Li, Z. X., Lloyd, J., Lott, F., Madec, G., Mancip, M., Marchand, M., Masson, S., Meurdesoif, Y., Mignot, J., Musat, I., Parouty, S., Polcher, J., Rio, C., Schulz, M., Swingedouw, D., Szopa, S., Talandier, C., Terray, P., Viovy, N., and

Vuichard, N.: Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5, Clim Dynam, doi: 10.1007/s00382-012-1636-1, 2013. 1-43, 2013.
Dümenil Gates, L., Hagemann, S., and Golz, C.: Observed historical discharge data from major rivers for climate model validation, Max Planck Institute for Meteorology, Report 307, 2000.
Dunne, J. P., John, J. G., Adcroft, A. J., Griffies, S. M., Hallberg, R. W., Shevliakova, E., Stouffer, R. J., Cooke, W., Dunne, K. A., Harrison, M. J., Krasting, J. P., Malyshev, S. L., Milly, P. C. D., Phillipps, P. J., Sentman, L. T., Samuels, B. L., Spelman, M. J., Winton, M., Wittenberg, A. T., and Zadeh, N.: GFDL's ESM2 Global Coupled Climate-Carbon Earth System Models. Part I: Physical Formulation and Baseline Simulation Characteristics, J Climate, 25, 6646-6665, 2012.
Dunne, J. P., John, J. G., Shevliakova, E., Stouffer, R. J., Krasting, J. P., Malyshev, S. L., Milly, P. C. D., Sentman, L. T., Adcroft, A. J., Cooke, W., Dunne, K. A., Griffies, S. M., Hallberg, R. W., Harrison, M. J., Levy, H., Wittenberg, A. T., Phillips, P. J., and Zadeh, N.: GFDL's ESM2 Global Coupled Climate-Carbon Earth System Models. Part II: Carbon System Formulation and Baseline Simulation Characteristics, J Climate, 26, 2247-2267, 2013.
Edgerton, E., Lavery, T., Hodges, M., and Bowser, J.: National dry deposition network: Second annual progress report, Tech. rep, 1990.
Emmons, L. K., Hauglustaine, D. A., Müller, J.-F., Carroll, M. A., Brasseur, G. P., Brunner, D., Staehelin, J., Thouret, V., and Marenco, A.: Data composites of airborne observations of tropospheric ozone and its precursors, J. Geophys. Res., 105, 20497-20538, 2000.
Eyring, V., Arblaster, J. M., Cionni, I., Sedlacek, J., Perliwitz, J., Young, P. J., Bekki, S., Bergmann, D., Cameron-Smith, P., Collins, W. J., Faluvegi, G., Gottschaldt, K. D., Horowitz, L. W., Kinnison, D. E., Lamarque, J. F., Marsh, D. R., Saint-Martin, D., Shindell, D. T., Sudo, K., Szopa, S., and Watanabe, S.: Long-term ozone changes and associated climate impacts in CMIP5 simulations, J Geophys Res-Atmos, 118, 5029-5060, 2013.
Eyring, V., Bony, S., Meehl, G. A., Senior, C., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organisation, Geosci. Model Dev. Discuss., 8, 10539-10583, 2015.
Eyring, V., Butchart, N., Waugh, D. W., Akiyoshi, H., Austin, J., Bekki, S., Bodeker, G. E., Boville, B. A., Brühl, C., Chipperfield, M. P., Cordero, E., Dameris, M., Deushi, M., Fioletov, V. E., Frith, S. M., Garcia, R. R., Gettelman, A., Giorgetta, M. A., Grewe, V., Jourdain, L., Kinnison, D. E., Mancini, E., Manzini, E., Marchand, M., Marsh, D. R., Nagashima, T., Newman, P. A., Nielsen, J. E., Pawson, S., Pitari, G., Plummer, D. A., Rozanov, E., Schraner, M., Shepherd, T. G., Shibata, K., Stolarski, R. S., Struthers, H., Tian, W., and Yoshiki, M.: Assessment of temperature, trace species, and ozone in chemistry-climate model simulations of the recent past, J. Geophys. Res., 111, D22308, 2006.
Eyring, V., Cionni, I., Bodeker, G. E., Charlton-Perez, A. J., Kinnison, D. E., Scinocca, J. F., Waugh, D. W., Akiyoshi, H., Bekki, S., Chipperfield, M. P., Dameris, M., Dhomse, S., Frith, S. M., Garny, H., Gettelman, A., Kubin, A., Langematz, U., Mancini, E., Marchand, M., Nakamura, T., Oman, L. D., Pawson, S., Pitari, G., Plummer, D. A., Rozanov, E., Shepherd, T. G., Shibata, K., Tian, W., Braesicke, P., Hardiman, S. C., Lamarque, J. F., Morgenstern, O., Pyle, J. A., Smale, D., and Yamashita, Y.: Multi-model assessment of stratospheric ozone return dates and ozone recovery in CCMVal-2 models, Atmos. Chem. Phys., 10, 9451-9472, 2010.
Feng, J., Liu, P., Chen, W., and Wang, X. C.: Contrasting Madden-Julian Oscillation activity during various stages of EP and CP El Ninos, Atmos Sci Lett, 16, 32-37, 2015.
Ferraro, R., Waliser, D. E., Gleckler, P., Taylor, K. E., and Eyring, V.: Evolving obs4MIPs to Support the Sixth Coupled Model Intercomparison Project (CMIP6), B Am Meteorol Soc, doi: 10.1175/BAMS-D-14-00216.1, 2015. 2015.

Fioletov, V. E., Bodeker, G. E., Miller, A. J., McPeters, R. D., and Stolarski, R.: Global and zonal total ozone variations estimated from ground-based and satellite measurements: 1964-2000, J Geophys Res-Atmos, 107, 2002.
Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., and Rummukainen, M.: Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Stocker, T. F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (Ed.), Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.
Free, M., Seidel, D. J., Angell, J. K., Lanzante, J., Durre, I., and Peterson, T. C.: Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC): A new data set of large-area anomaly time series, J Geophys Res-Atmos, 110, 2005.
Friedlingstein, P., Cox, P., Betts, R., Bopp, L., von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H. D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K.-G., Schnur, R., Strassmann, K., Weaver, A. J., Yoshikawa, C., and Zeng, N.: Climate–Carbon Cycle Feedback Analysis: Results from the C4MIP Model Intercomparison, J Climate, 19, 3337-3353, 2006.
Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K., and Knutti, R.: Uncertainties in CMIP5 Climate Projections due to Carbon Cycle Feedbacks, J Climate, 27, 511-526, 2014.
Frolicher, T. L., Sarmiento, J. L., Paynter, D. J., Dunne, J. P., Krasting, J. P., and Winton, M.: Dominance of the Southern Ocean in Anthropogenic Carbon and Heat Uptake in CMIP5 Models, J Climate, 28, 862-886, 2015.
GCOS: Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC, August 2010, 2010. 2010.
Gettelman, A., Eyring, V., Fischer, C., Shiona, H., Cionni, I., Neish, M., Morgenstern, O., Wood, S. W., and Li, Z.: A community diagnostic tool for chemistry climate model validation, Geosci. Model Dev., 5, 1061-1073, 2012.
GEWEX-news, Vol. 21, No. 1, February 2011.
Ghan, S. J. and Schwartz, S. E.: Aerosol properties and processes - A path from field and laboratory measurements to global climate models, B Am Meteorol Soc, 88, 1059-+, 2007.
Gibbs, H. K.: Olson's Major World Ecosystem Complexes Ranked by Carbon in Live Vegetation: An Updated Database Using the GLC2000 Land Cover Product (NDP-017b), doi: DOI: 10.3334/CDIAC/lue.ndp017.2006, 2006. 2006.
Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Bottinger, M., Brovkin, V., Crueger, T., Esch, M., Fieg, K., Glushak, K., Gayler, V., Haak, H., Hollweg, H. D., Ilyina, T., Kinne, S., Kornblueh, L., Matei, D., Mauritsen, T., Mikolajewicz, U., Mueller, W., Notz, D., Pithan, F., Raddatz, T., Rast, S., Redler, R., Roeckner, E., Schmidt, H., Schnur, R., Segschneider, J., Six, K. D., Stockhause, M., Timmreck, C., Wegner, J., Widmann, H., Wieners, K. H., Claussen, M., Marotzke, J., and Stevens, B.: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5, Journal of Advances in Modeling Earth Systems, 5, 572-597, 2013.
Gleckler, P. J., Doutriaux, C., Durack P. J., Taylor K. E. , Zhang, Y., Williams, D. N., Mason, E., and Servonnat, J.: A More Powerful Reality Test for Climate Models, Eos Trans. AGU, in press, 2016.
Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res., 113, D06104, 2008.

GLOBALVIEW-CO2: Cooperative Atmospheric Data Integration Project - Carbon Dioxide, CD-ROM, NOAA ESRL, Boulder, Colorado, 2008.

Goswami, B. N., Krishnamurthy, V., and Annamalai, H.: A broad-scale circulation index for the interannual variability of the Indian summer monsoon, Q J Roy Meteor Soc, 125, 611-633, 1999.

Grooß, J. U. and Russell Iii, J. M.: Technical note: A stratospheric climatology for O_3, H_2O, CH_4, NOx, HCl and HF derived from HALOE measurements, Atmos. Chem. Phys., 5, 2797-2807, 2005.

Guilyardi, E.: El Nino-mean state-seasonal cycle interactions in a multi-model ensemble, Clim Dynam, 26, 329-348, 2006.

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y. H., Ciais, P., Fan, S. M., Fung, I. Y., Gloor, M., Heimann, M., Higuchi, K., John, J., Kowalczyk, E., Maki, T., Maksyutov, S., Peylin, P., Prather, M., Pak, B. C., Sarmiento, J., Taguchi, S., Takahashi, T., and Yuen, C. W.: TransCom 3 CO_2 inversion intercomparison: 1. Annual mean control results and sensitivity to transport and prior flux information, Tellus B, 55, 555-579, 2003.

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Pak, B. C., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y. H., Ciais, P., Fung, I. Y., Heimann, M., John, J., Maki, T., Maksyutov, S., Peylin, P., Prather, M., and Taguchi, S.: Transcom 3 inversion intercomparison: Model mean results for the estimation of seasonal carbon sources and sinks, Global Biogeochem Cy, 18, 2004.

Hagemann, S., Loew, A., and Andersson, A.: Combined evaluation of MPI-ESM land surface water and energy fluxes, Journal of Advances in Modeling Earth Systems, 5, 259-286, 2013.

Hagemann, S., Machenhauer, B., Jones, R., Christensen, O. B., Deque, M., Jacob, D., and Vidale, P. L.: Evaluation of water and energy budgets in regional climate models applied over Europe, Clim Dynam, 23, 547-567, 2004.

Hall, A. and Qu, X.: Using the current seasonal cycle to constrain snow albedo feedback in future climate change, Geophys Res Lett, 33, 2006.

Hazeleger, W., Wang, X., Severijns, C., Stefanescu, S., Bintanja, R., Sterl, A., Wyser, K., Semmler, T., Yang, S., van den Hurk, B., van Noije, T., van der Linden, E., and van der Wiel, K.: EC-Earth V2.2: description and validation of a new seamless earth system prediction model, Clim Dynam, 39, 2611-2629, 2012.

Held, I. M., Delworth, T. L., Lu, J., Findell, K. L., and Knutson, T. R.: Simulation of Sahel drought in the 20th and 21st centuries, P Natl Acad Sci USA, 102, 17891-17896, 2005.

Hoell, A., Barlow, M., Wheeler, M. C., and Funk, C.: Disruptions of El Niño–Southern Oscillation Teleconnections by the Madden–Julian Oscillation, Geophys Res Lett, 41, 998-1004, 2014.

Holben, B. N., Eck, T. F., Slutsker, I., Tanre, D., Buis, J. P., Setzer, A., Vermote, E., Reagan, J. A., Kaufman, Y. J., Nakajima, T., Lavenu, F., Jankowiak, I., and Smirnov, A.: AERONET - A federated instrument network and data archive for aerosol characterization, Remote Sens Environ, 66, 1-16, 1998.

Huffman, G. J., Adler, R. F., Bolvin, D. T., Gu, G. J., Nelkin, E. J., Bowman, K. P., Hong, Y., Stocker, E. F., and Wolff, D. B.: The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales, J Hydrometeorol, 8, 38-55, 2007.

Huffman, G. J., Adler, R. F., Morrissey, M. M., Bolvin, D. T., Curtis, S., Joyce, R., McGavock, B., and Susskind, J.: Global Precipitation at One-Degree Daily Resolution from Multisatellite Observations, J Hydrometeorol, 2, 36-50, 2001.

Hung, M. P., Lin, J. L., Wang, W. Q., Kim, D., Shinoda, T., and Weaver, S. J.: MJO and Convectively Coupled Equatorial Waves Simulated by CMIP5 Climate Models, J Climate, 26, 6185-6214, 2013.

Hurrell, J. W. and Deser, C.: North Atlantic climate variability: The role of the North Atlantic Oscillation, J Marine Syst, 78, 28-41, 2009.

Iguchi, T.: Correlations between interannual variations of simulated global and regional CO2 fluxes from terrestrial ecosystems and El Nino Southern Oscillation, Tellus B, 63, 196-204, 2011.
Ihaka, R. and Gentleman, R.: R: A Language for Data Analysis and Graphics, Journal of Computational and Graphical Statistics, 5, 299-314, 1996.
Ilyina, T., Six, K. D., Segschneider, J., Maier-Reimer, E., Li, H. M., and Nunez-Riboni, I.: Global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-Earth system model in different CMIP5 experimental realizations, Journal of Advances in Modeling Earth Systems, 5, 287-315, 2013.
IPCC: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.
Jiang, J. H., Su, H., Zhai, C. X., Shen, T. J., Wu, T. W., Zhang, J., Cole, J. N. S., von Salzen, K., Donner, L. J., Seman, C., Del Genio, A., Nazarenko, L. S., Dufresne, J. L., Watanabe, M., Morcrette, C., Koshiro, T., Kawai, H., Gettelman, A., Millan, L., Read, W. G., Livesey, N. J., Kasai, Y., and Shiotani, M.: Evaluating the Diurnal Cycle of Upper-Tropospheric Ice Clouds in Climate Models Using SMILES Observations, J Atmos Sci, 72, 1022-1044, 2015.
Jöckel et al., P.: Earth System Chemistry Integrated Modelling (ESCiMo) with the Modular Earth Submodel System (MESSy, version 2.51), Geosci. Model Dev., submitted, 2015.
Jöckel, P., Kerkweg, A., Pozzer, A., Sander, R., Tost, H., Riede, H., Baumgaertner, A., Gromov, S., and Kern, B.: Development cycle 2 of the Modular Earth Submodel System (MESSy2), Geosci Model Dev, 3, 717-752, 2010.
Jones, C. D., Collins, M., Cox, P. M., and Spall, S. A.: The carbon cycle response to ENSO: A coupled climate-carbon cycle model study, J Climate, 14, 4113-4129, 2001.
Jung, M., Reichstein, M., and Bondeau, A.: Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model, Biogeosciences, 6, 2001-2013, 2009.
Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S. K., Hnilo, J. J., Fiorino, M., and Potter, G. L.: Ncep-Doe Amip-Ii Reanalysis (R-2), B Am Meteorol Soc, 83, 1631-1643, 2002.
Kim, D., Sperber, K., Stern, W., Waliser, D., Kang, I. S., Maloney, E., Wang, W., Weickmann, K., Benedict, J., Khairoutdinov, M., Lee, M. I., Neale, R., Suarez, M., Thayer-Calder, K., and Zhang, G.: Application of MJO Simulation Diagnostics to Climate Models, J Climate, 22, 6413-6436, 2009.
King, M. D., Menzel, W. P., Kaufman, Y. J., Tanre, D., Gao, B. C., Platnick, S., Ackerman, S. A., Remer, L. A., Pincus, R., and Hubanks, P. A.: Cloud and aerosol properties, precipitable water, and profiles of temperature and water vapor from MODIS, Ieee T Geosci Remote, 41, 442-458, 2003.
Kinne, S., Schulz, M., Litvinov, P., Stebel, K., Holzer-Popp, T., and de Leeuw, G.: ATSR Climate Data Record Evaluation Report, version 1.2, ESA, Aerosol_cci,, http://www.esa-aerosol-cci.org/?q=webfm_send/836, 2015.
Kinne, S., Schulz, M., Textor, C., Guibert, S., Balkanski, Y., Bauer, S. E., Berntsen, T., Berglen, T. F., Boucher, O., Chin, M., Collins, W., Dentener, F., Diehl, T., Easter, R., Feichter, J., Fillmore, D., Ghan, S., Ginoux, P., Gong, S., Grini, A., Hendricks, J. E., Herzog, M., Horowitz, L., Isaksen, I., Iversen, T., Kirkavag, A., Kloster, S., Koch, D., Kristjansson, J. E., Krol, M., Lauer, A., Lamarque, J. F., Lesins, G., Liu, X., Lohmann, U., Montanaro, V., Myhre, G., Penner, J. E., Pitari, G., Reddy, S., Seland, O., Stier, P., Takemura, T., and Tie, X.: An AeroCom initial assessment - optical properties in aerosol component modules of global models, Atmos Chem Phys, 6, 1815-1834, 2006.
Kistler, R., Collins, W., Saha, S., White, G., Woollen, J., Kalnay, E., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., van den Dool, H., Jenne, R., and Fiorino, M.: The NCEP–NCAR 50–Year Reanalysis: Monthly Means CD–ROM and Documentation, B Am Meteorol Soc, 82, 247-267, 2001.

Klein, S. A., Zhang, Y. Y., Zelinka, M. D., Pincus, R., Boyle, J., and Gleckler, P. J.: Are climate model simulations of clouds improving? An evaluation using the ISCCP simulator, J Geophys Res-Atmos, 118, 1329-1342, 2013.
Klotzbach, P. J.: The Madden-Julian Oscillation's Impacts on Worldwide Tropical Cyclone Activity, J Climate, 27, 2317-2330, 2014.
Krishnamurthy, V. and Misra, V.: Daily atmospheric variability in the South American monsoon system, Clim Dynam, 37, 803-819, 2011.
Landschützer, P., Gruber, N., Bakker, D. C. E., and Schuster, U.: An observation-based global monthly gridded sea surface pCO2 product from 1998 through 2011 and its monthly climatology. Carbon Dioxide Information Analysis Center, O. R. N. L., US Department of Energy (Ed.), Oak Ridge, Tennessee, 2014a.
Landschützer, P., Gruber, N., Bakker, D. C. E., and Schuster, U.: Recent variability of the global ocean carbon sink, Global Biogeochem Cy, 28, 927-949, 2014b.
Lauer, A. and Hamilton, K.: Simulating Clouds with Global Climate Models: A Comparison of CMIP5 Results with CMIP3 and Satellite Data, J Climate, 26, 3823-3845, 2013.
Lauer, A., Hendricks, J., Ackermann, I., Schell, B., Hass, H., and Metzger, S.: Simulating aerosol microphysics with the ECHAM/MADE GCM - Part I: Model description and comparison with observations, Atmos Chem Phys, 5, 3251-3276, 2005.
Le Quéré, C., Moriarty, R., Andrew, R. M., Peters, G. P., Ciais, P., Friedlingstein, P., Jones, S. D., Sitch, S., Tans, P., Arneth, A., Boden, T. A., Bopp, L., Bozec, Y., Canadell, J. G., Chevallier, F., Cosca, C. E., Harris, I., Hoppema, M., Houghton, R. A., House, J. I., Jain, A., Johannessen, T., Kato, E., Keeling, R. F., Kitidis, V., Klein Goldewijk, K., Koven, C., Landa, C. S., Landschützer, P., Lenton, A., Lima, I. D., Marland, G., Mathis, J. T., Metzl, N., Nojiri, Y., Olsen, A., Ono, T., Peters, W., Pfeil, B., Poulter, B., Raupach, M. R., Regnier, P., Rödenbeck, C., Saito, S., Salisbury, J. E., Schuster, U., Schwinger, J., Séférian, R., Segschneider, J., Steinhoff, T., Stocker, B. D., Sutton, A. J., Takahashi, T., Tilbrook, B., van der Werf, G. R., Viovy, N., Wang, Y. P., Wanninkhof, R., Wiltshire, A., and Zeng, N.: Global carbon budget 2014, Earth Syst. Sci. Data Discuss., 7, 521-610, 2014.
Lee, Y. H., Lamarque, J. F., Flanner, M. G., Jiao, C., Shindell, D. T., Berntsen, T., Bisiaux, M. M., Cao, J., Collins, W. J., Curran, M., Edwards, R., Faluvegi, G., Ghan, S., Horowitz, L. W., McConnell, J. R., Ming, J., Myhre, G., Nagashima, T., Naik, V., Rumbold, S. T., Skeie, R. B., Sudo, K., Takemura, T., Thevenon, F., Xu, B., and Yoon, J. H.: Evaluation of preindustrial to present-day black carbon and its albedo forcing from Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP), Atmos Chem Phys, 13, 2607-2634, 2013.
Legates, D. R. and Willmott, C. J.: Mean seasonal and spatial variability in gauge-corrected, global precipitation, International Journal of Climatology, 10, 111-127, 1990.
Levine, R. C., Turner, A. G., Marathayil, D., and Martin, G. M.: The role of northern Arabian Sea surface temperature biases in CMIP5 model simulations and future projections of Indian summer monsoon rainfall, Clim Dynam, 41, 155-172, 2013.
Li, G. and Xie, S. P.: Tropical Biases in CMIP5 Multimodel Ensemble: The Excessive Equatorial Pacific Cold Tongue and Double ITCZ Problems, J Climate, 27, 1765-1780, 2014.
Liebmann, B. and Smith, C. A.: Description of a complete (interpolated) outgoing longwave radiation dataset, B Am Meteorol Soc, 77, 1275-1277, 1996.
Lin, J. L.: The double-ITCZ problem in IPCC AR4 coupled GCMs: Ocean-atmosphere feedback analysis, J Climate, 20, 4497-4525, 2007.
Lin, J. L., Kiladis, G. N., Mapes, B. E., Weickmann, K. M., Sperber, K. R., Lin, W., Wheeler, M. C., Schubert, S. D., Del Genio, A., Donner, L. J., Emori, S., Gueremy, J. F., Hourdin, F., Rasch, P. J., Roeckner, E., and Scinocca, J. F.: Tropical intraseasonal variability in 14 IPCC AR4 climate models. Part I: Convective signals, J Climate, 19, 2665-2690, 2006.

Lin, J. L., Weickman, K. M., Kiladis, G. N., Mapes, B. E., Schubert, S. D., Suarez, M. J., Bacmeister, J. T., and Lee, M. I.: Subseasonal variability associated with Asian summer monsoon simulated by 14 IPCC AR4 coupled GCMs, J Climate, 21, 4541-4567, 2008.

Locarnini, R. A., Mishonov, A. V., Antonov, J. I., Boyer, T. P., Garcia, H. E., Baranova, O. K., Zweng, M. M., and Johnson, D. R.: World Ocean Atlas 2009, Volume 1: Temperature. In: NOAA Atlas NESDIS 68,, Levitus, S. (Ed.), U.S. Government Printing Office, Washington, D.C., 2010.

Loeb, N. G., Lyman, J. M., Johnson, G. C., Allan, R. P., Doelling, D. R., Wong, T., Soden, B. J., and Stephens, G. L.: Observed changes in top-of-the-atmosphere radiation and upper-ocean heating consistent within uncertainty, Nat Geosci, 5, 110-113, 2012.

Loeb, N. G., Wielicki, B. A., Doelling, D. R., Smith, G. L., Keyes, D. F., Kato, S., Manalo-Smith, N., and Wong, T.: Toward Optimal Closure of the Earth's Top-of-Atmosphere Radiation Budget, J Climate, 22, 748-766, 2009.

Lohmann, U. and Feichter, J.: Global indirect aerosol effects: a review, Atmos. Chem. Phys., 5, 715-737, 2005.

Loyola, D. G. and Coldewey-Egbers, M.: Multi-sensor data merging with stacked neural networks for the creation of satellite long-term climate data records, Eurasip J Adv Sig Pr, doi: 10.1186/1687-6180-2012-91, 2012. 2012.

Mace, G. G.: Cloud properties and radiative forcing over the maritime storm tracks of the Southern Ocean and North Atlantic derived from A-Train, J Geophys Res-Atmos, 115, 2010.

Madden, R. A. and Julian, P. R.: Detection of a 40–50 Day Oscillation in the Zonal Wind in the Tropical Pacific, J. Atmos. Sci., 28, 702–708, 1971.

Madec, G.: NEMO ocean engine. Note du Pole de modélisation, Institut Pierre-Simon Laplace (IPSL), France, No 27, ISSN No 1288-1619, 2008.

Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., and Francis, R. C.: A Pacific interdecadal climate oscillation with impacts on salmon production, B Am Meteorol Soc, 78, 1069-1079, 1997.

Mathon, V., Laurent, H., and Lebel, T.: Mesoscale convective system rainfall in the Sahel, J Appl Meteorol, 41, 1081-1092, 2002.

McClain, C. R., Cleave, M. L., Feldman, G. C., Gregg, W. W., Hooker, S. B., and Kuring, N.: Science quality SeaWiFS data for global biosphere research, Sea Technol, 39, 10-16, 1998.

Meehl, G. A., Moss, R., Taylor, K. E., Eyring, V., Stouffer, R. J., Bony, S., and Stevens, B.: Climate Model Intercomparisons: Preparing for the Next Phase, Eos Trans. AGU, 59, 77, 2014.

Meier, W., Fetterer, F., Savoie, M., Mallory, S., Duerr, R., and Stroeve, J.: NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration. Version 2. [sea ice concentration]. Center, N. S. a. I. D. (Ed.), Boulder, Colorado USA, 2013.

Miller, A. J., Nagatani, R. M., Flynn, L. E., Kondragunta, S., Beach, E., Stolarski, R., McPeters, R. D., Bhartia, P. K., DeLand, M. T., Jackman, C. H., Wuebbles, D. J., Patten, K. O., and Cebula, R. P.: A cohesive total ozone data set from the SBUV(/2) satellite system, J Geophys Res-Atmos, 107, 2002.

Mitchell, T. D. and Jones, P. D.: An improved method of constructing a database of monthly climate observations and associated high-resolution grids, International Journal of Climatology, 25, 693-712, 2005.

Mueller, B., Hirschi, M., Jimenez, C., Ciais, P., Dirmeyer, P. A., Dolman, A. J., Fisher, J. B., Jung, M., Ludwig, F., Maignan, F., Miralles, D. G., McCabe, M. F., Reichstein, M., Sheffield, J., Wang, K., Wood, E. F., Zhang, Y., and Seneviratne, S. I.: Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis, Hydrol. Earth Syst. Sci., 17, 3707-3720, 2013.

Mueller, B. and Seneviratne, S. I.: Systematic land climate and evapotranspiration biases in CMIP5 simulations, Geophys Res Lett, 41, 128-134, 2014.

Myhre, G., D. Shindell, F.-M. Bréon, W. Collins, J. Fuglestvedt, J. Huang, D. Koch, J.-F. Lamarque, D. Lee, B. Mendoza, T. Nakajima, A. Robock, G. Stephens, Takemura, T., and Zhang, H.: Anthropogenic and Natural Radiative Forcing. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Stocker, T. F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (Ed.), Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Nachtergaele, F., H. v. V., L. Verekst, and Widberg, D.: Harmonized World Soil Database v 1.2., 2012.

Nam, C., Bony, S., Dufresne, J. L., and Chepfer, H.: The 'too few, too bright' tropical low-cloud problem in CMIP5 models, Geophys Res Lett, 39, 2012.

NCL: The NCAR Command Language (Version 6.3.0) [Software], Boulder, Colorado, UCAR/NCAR/CISL/TDD, http://dx.doi.org/10.5065/D6WD3XH5. 2016.

Nicholson, S. E., Some, B., and Kone, B.: An analysis of recent rainfall conditions in West Africa, including the rainy seasons of the 1997 El Nino and the 1998 La Nina years, J Climate, 13, 2628-2640, 2000.

Notz, D., Haumann, F. A., Haak, H., Jungclaus, J. H., and Marotzke, J.: Arctic sea-ice evolution as modeled by Max Planck Institute for Meteorology's Earth system model, Journal of Advances in Modeling Earth Systems, doi: 10.1002/jame.20016, 2013. n/a-n/a, 2013.

O'Dell, C. W., Wentz, F. J., and Bennartz, R.: Cloud liquid water path from satellite-based passive microwave observations: A new climatology over the global oceans, J Climate, 21, 1721-1739, 2008.

Olson, J. S., Watts, J. A., and Allison, L. J.: Major world ecosystem complexes ranked by carbon in live vegetation: A database (NDP-017). Carbon Dioxide Information Analysis Center. 1985.

Orlowsky, B. and Seneviratne, S. I.: Elusive drought: uncertainty in observed trends and short- and long-term CMIP5 projections, Hydrol Earth Syst Sc, 17, 1765-1781, 2013.

Oueslati, B. and Bellon, G.: The double ITCZ bias in CMIP5 models: interaction between SST, large-scale circulation and precipitation, Clim Dynam, 44, 585-607, 2015.

Pai, D. S., Bhate, J., Sreejith, O. P., and Hatwar, H. R.: Impact of MJO on the intraseasonal variation of summer monsoon rainfall over India, Clim Dynam, 36, 41-55, 2011.

Peng, G., Meier, W. N., Scott, D. J., and Savoie, M. H.: A long-term and reproducible passive microwave sea ice concentration data record for climate studies and monitoring, Earth Syst. Sci. Data, 5, 311-318, 2013.

Phillips, A. S., Deser, C., and Fasullo, J.: Evaluating Modes of Variability in Climate Models, Eos Trans. AGU, 95(49), 453-455, 2014.

Pierrehumbert, R. T.: Thermostats, Radiator Fins, and the Local Runaway Greenhouse, J Atmos Sci, 52, 1784-1806, 1995.

Pincus, R., Batstone, C. P., Hofmann, R. J. P., Taylor, K. E., and Glecker, P. J.: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models, J. Geophys. Res., 113, D14209, 2008.

Pincus, R., Platnick, S., Ackerman, S. A., Hemler, R. S., and Hofmann, R. J. P.: Reconciling Simulated and Observed Views of Clouds: MODIS, ISCCP, and the Limits of Instrument Simulators, J Climate, 25, 4699-4720, 2012.

Pozzer, A., de Meij, A., Pringle, K. J., Tost, H., Doering, U. M., van Aardenne, J., and Lelieveld, J.: Distributions and regional budgets of aerosols and their precursors simulated with the EMAC chemistry-climate model, Atmos Chem Phys, 12, 961-987, 2012.

Pringle, K. J., Tost, H., Message, S., Steil, B., Giannadaki, D., Nenes, A., Fountoukis, C., Stier, P., Vignati, E., and Leieved, J.: Description and evaluation of GMXe: a new aerosol submodel for global simulations (v1), Geosci Model Dev, 3, 391-412, 2010.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, J. Geophys. Res., 108, 4407, 2003.

Redelsperger, J. L., Thorncroft, C. D., Diedhiou, A., Lebel, T., Parker, D. J., and Polcher, J.: African monsoon multidisciplinary analysis - An international research project and field campaign, B Am Meteorol Soc, 87, 1739-+, 2006.

Reichler, T. and Kim, J.: How Well Do Coupled Models Simulate Today's Climate?, B Am Meteorol Soc, 89, 303-311, 2008.

Richter, I., Behera, S. K., Doi, T., Taguchi, B., Masumoto, Y., and Xie, S. P.: What controls equatorial Atlantic winds in boreal spring?, Clim Dynam, 43, 3091-3104, 2014.

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G. K., Bloom, S., Chen, J. Y., Collins, D., Conaty, A., Da Silva, A., Gu, W., Joiner, J., Koster, R. D., Lucchesi, R., Molod, A., Owens, T., Pawson, S., Pegion, P., Redder, C. R., Reichle, R., Robertson, F. R., Ruddick, A. G., Sienkiewicz, M., and Woollen, J.: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications, J Climate, 24, 3624-3648, 2011.

Righi, M., Eyring, V., Gottschaldt, K. D., Klinger, C., Frank, F., Jöckel, P., and Cionni, I.: Quantitative evaluation of ozone and selected climate parameters in a set of EMAC simulations, Geosci. Model Dev., 8, 733-768, 2015.

Righi, M., Hendricks, J., and Sausen, R.: The global impact of the transport sectors on atmospheric aerosol: simulations for year 2000 emissions, Atmos Chem Phys, 13, 9939-9970, 2013.

Rio, C., Hourdin, F., Grandpeix, J. Y., and Lafore, J. P.: Shifting the diurnal cycle of parameterized deep convection over land, Geophys Res Lett, 36, 2009.

Rodenbeck, C., Bakker, D. C. E., Metzl, N., Olsen, A., Sabine, C., Cassar, N., Reum, F., Keeling, R. F., and Heimann, M.: Interannual sea-air CO2 flux variability from an observation-driven ocean mixed-layer scheme, Biogeosciences, 11, 4599-4613, 2014.

Roehrig, R., Bouniol, D., Guichard, F., Hourdin, F., and Redelsperger, J. L.: The Present and Future of the West African Monsoon: A Process-Oriented Assessment of CMIP5 Simulations along the AMMA Transect, J Climate, 26, 6471-6505, 2013.

Rossow, W. B. and Schiffer, R. A.: Advances in Understanding Clouds from ISCCP, B Am Meteorol Soc, 80, 2261-2287, 1999.

Rossow, W. B. and Schiffer, R. A.: ISCCP Cloud Data Products, B Am Meteorol Soc, 72, 2-20, 1991.

Russell, J. M., III, Gordley, L. L., Park, J. H., Drayson, S. R., Hesketh, W. D., Cicerone, R. J., Tuck, A. F., Frederick, J. E., Harries, J. E., and Crutzen, P. J.: THE HALOGEN OCCULTATION EXPERIMENT, J. Geophys. Res., 98, 10777-10797, 1993.

Sabeerali, C. T., Dandi, A., Dhakate, A., Salunke, K., Mahapatra, S., and Rao, S. A.: Simulation of boreal summer intraseasonal oscillations in the latest CMIP5 coupled GCMs, J Geophys Res-Atmos, 118, 4401-4420, 2013.

Sanderson, B. M., Knutti, R., and Caldwell, P.: Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties, J Climate, 28, 5150-5170, 2015a.

Sanderson, B. M., Knutti, R., and Caldwell, P.: A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble, J Climate, 28, 5171-5194, 2015b.

Schulz, M., Textor, C., Kinne, S., Balkanski, Y., Bauer, S., Berntsen, T., Berglen, T., Boucher, O., Dentener, F., Guibert, S., Isaksen, I. S. A., Iversen, T., Koch, D., Kirkevag, A., Liu, X., Montanaro, V., Myhre, G., Penner, J. E., Pitari, G., Reddy, S., Seland, O., Stier, P., and Takemura, T.: Radiative forcing by aerosols as derived from the AeroCom present-day and pre-industrial simulations, Atmos Chem Phys, 6, 5225-5246, 2006.

Shi, Y., Zhang, J., Reid, J. S., Holben, B., Hyer, E. J., and Curtis, C.: An analysis of the collection 5 MODIS over-ocean aerosol optical depth product for its implication in aerosol assimilation, Atmos Chem Phys, 11, 557-565, 2011.

Smith, G. L., Mlynczak, P. E., Rutan, D. A., and Wong, T.: Comparison of the Diurnal Cycle of Outgoing Longwave Radiation from a Climate Model with Results from ERBE, J Appl Meteorol Clim, 47, 3188-3201, 2008.

SPARC-CCMVal: SPARC Report on the Evaluation of Chemistry-Climate Models, V. Eyring, T. G. Shepherd, D. W. Waugh (Eds.). SPARC Report No. 5, WCRP-132, WMO/TD-No. 1526., 2010.

Sperber, K., Annamalai, H., Kang, I. S., Kitoh, A., Moise, A., Turner, A., Wang, B., and Zhou, T.: The Asian summer monsoon: an intercomparison of CMIP5 vs. CMIP3 simulations of the late 20th century, Clim Dynam, 41, 2711-2744, 2013.

Stephens, G. L. and Greenwald, T. J.: The Earths Radiation Budget and Its Relation to Atmospheric Hydrology .1. Observations of the Clear Sky Greenhouse-Effect, J Geophys Res-Atmos, 96, 15311-15324, 1991.

Stephens, G. L., Vane, D. G., Boain, R. J., Mace, G. G., Sassen, K., Wang, Z. E., Illingworth, A. J., O'Connor, E. J., Rossow, W. B., Durden, S. L., Miller, S. D., Austin, R. T., Benedetti, A., Mitrescu, C., and Team, C. S.: The cloudsat mission and the a-train - A new dimension of space-based observations of clouds and precipitation, B Am Meteorol Soc, 83, 1771-1790, 2002.

Sterl, A., Bintanja, R., Brodeau, L., Gleeson, E., Koenigk, T., Schmith, T., Semmler, T., Severijns, C., Wyser, K., and Yang, S. T.: A look at the ocean in the EC-Earth climate model, Clim Dynam, 39, 2631-2657, 2012.

Stevens, B. and Schwartz, S. E.: Observing and Modeling Earth's Energy Flows, Surv Geophys, 33, 779-816, 2012.

Stolarski, R. S. and Frith, S. M.: Search for evidence of trend slow-down in the long-term TOMS/SBUV total ozone data record: the importance of instrument drift uncertainty, Atmos Chem Phys, 6, 4057-4065, 2006.

Stowasser, M., Annamalai, H., and Hafner, J.: Response of the South Asian Summer Monsoon to Global Warming: Mean and Synoptic Systems, J Climate, 22, 1014-1036, 2009.

Stroeve, J., Holland, M. M., Meier, W., Scambos, T., and Serreze, M.: Arctic sea ice decline: Faster than forecast, Geophys. Res. Lett., 34, L09501, 2007.

Stroeve, J. C., Kattsov, V., Barrett, A., Serreze, M., Pavlova, T., Holland, M., and Meier, W. N.: Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations, Geophys Res Lett, 39, 2012.

Takahashi, T., Sutherland, S. C., Chipman, D. W., Goddard, J. G., Ho, C., Newberger, T., Sweeney, C., and Munro, D. R.: Climatological distributions of pH, pCO2, total CO2, alkalinity, and CaCO3 saturation in the global surface ocean, and temporal changes at selected locations, Mar. Chem., 164, 95–125, 2014.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram., J Geophys Res-Atmos, 106, 7183-7192, 2001.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of Cmip5 and the Experiment Design, B Am Meteorol Soc, 93, 485-498, 2012.

Teixeira, J., Waliser, D., Ferraro, R., Gleckler, P., Lee, T., and Potter, G.: Satellite Observations for CMIP5: The Genesis of Obs4MIPs, B Am Meteorol Soc, 95, 1329-1334, 2014.

Thompson, D. W. J. and Wallace, J. M.: Annular modes in the extratropical circulation. Part I: Month-to-month variability, J Climate, 13, 1000-1016, 2000.

Tilmes, S., Lamarque, J. F., Emmons, L. K., Conley, A., Schultz, M. G., Saunois, M., Thouret, V., Thompson, A. M., Oltmans, S. J., Johnson, B., and Tarasick, D.: Ozonesonde climatology between 1995 and 2009: description, evaluation and applications, Atmos. Chem. Phys. Discuss., 11, 28747-28796, 2011.

Totsuka, T., Sase, H., and Shimizu, H.: Major activities of acid deposition monitoring network in East Asia (EANET) and related studies. In: Plant Responses to Air Pollution and Global Change, Omasa, K., Nouchi, I., and De Kok, L. (Eds.), Springer Japan, 2005.
Trenberth, K. E. and Fasullo, J. T.: An observational estimate of inferred ocean energy divergence, J Phys Oceanogr, 38, 984-999, 2008.
Trenberth, K. E. and Fasullo, J. T.: Simulation of Present-Day and Twenty-First-Century Energy Budgets of the Southern Oceans, J Climate, 23, 440-454, 2010.
Trenberth, K. E. and Shea, D. J.: Atlantic hurricanes and natural variability in 2005, Geophys Res Lett, 33, 2006.
Turner, A. G., Inness, P. M., and Slingo, J. M.: The role of the basic state in the ENSO-monsoon relationship and implications for predictability, Q J Roy Meteor Soc, 131, 781-804, 2005.
Uppala, S. M., KÅllberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L. V. D., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., McNally, A. P., Mahfouf, J. F., Morcrette, J. J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J.: The ERA-40 re-analysis, Q J Roy Meteor Soc, 131, 2961-3012, 2005.
Voulgarakis, A., Naik, V., Lamarque, J. F., Shindell, D. T., Young, P. J., Prather, M. J., Wild, O., Field, R. D., Bergmann, D., Cameron-Smith, P., Cionni, I., Collins, W. J., Dalsoren, S. B., Doherty, R. M., Eyring, V., Faluvegi, G., Folberth, G. A., Horowitz, L. W., Josse, B., MacKenzie, I. A., Nagashima, T., Plummer, D. A., Righi, M., Rumbold, S. T., Stevenson, D. S., Strode, S. A., Sudo, K., Szopa, S., and Zeng, G.: Analysis of present day and future OH and methane lifetime in the ACCMIP simulations, Atmos Chem Phys, 13, 2563-2587, 2013.
Waliser, D., Sperber, K., Hendon, H., Kim, D., Wheeler, M., Weickmann, K., Zhang, C., Donner, L., Gottschalck, J., Higgins, W., Kang, I. S., Legler, D., Moncrieff, M., Vitart, F., Wang, B., Wang, W., Woolnough, S., Maloney, E., Schubert, S., Stern, W., and Oscillation, C. M.-J.: MJO Simulation Diagnostics, J Climate, 22, 3006-3030, 2009.
Wang, B. and Fan, Z.: Choice of south Asian summer monsoon indices, B Am Meteorol Soc, 80, 629-638, 1999.
Wang, B., Liu, J., Kim, H. J., Webster, P. J., and Yim, S. Y.: Recent change of the global monsoon precipitation (1979-2008), Clim Dynam, 39, 1123-1135, 2012.
Waugh, D. W. and Eyring, V.: Quantitative performance metrics for stratospheric-resolving chemistry-climate models, Atmos. Chem. Phys., 8, 5699-5713, 2008.
Webster, P. J. and Yang, S.: Monsoon and Enso - Selectively Interactive Systems, Q J Roy Meteor Soc, 118, 877-926, 1992.
Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resour Res, 50, 7505-7514, 2014.
Wenzel, S., Cox, P. M., Eyring, V., and Friedlingstein, P.: Emergent constraints on climate-carbon cycle feedbacks in the CMIP5 Earth system models, Journal of Geophysical Research: Biogeosciences, 119, 2013JG002591, 2014.
Wielicki, B. A., Barkstrom, B. R., Harrison, E. F., Lee, R. B., Louis Smith, G., and Cooper, J. E.: Clouds and the Earth's Radiant Energy System (CERES): An Earth Observing System Experiment, B Am Meteorol Soc, 77, 853-868, 1996.
Williams, K. and Webb, M.: A quantitative performance assessment of cloud regimes in climate models, Clim Dynam, 33, 141-157, 2009.

Xie, P. and Arkin, P. A.: Global Precipitation: A 17-Year Monthly Analysis Based on Gauge Observations, Satellite Estimates, and Numerical Model Outputs, B Am Meteorol Soc, 78, 2539-2558, 1997.
Young, P. J., Archibald, A. T., Bowman, K. W., Lamarque, J. F., Naik, V., Stevenson, D. S., Tilmes, S., Voulgarakis, A., Wild, O., Bergmann, D., Cameron-Smith, P., Cionni, I., Collins, W. J., Dalsøren, S. B., Doherty, R. M., Eyring, V., Faluvegi, G., Horowitz, L. W., Josse, B., Lee, Y. H., MacKenzie, I. A., Nagashima, T., Plummer, D. A., Righi, M., Rumbold, S. T., Skeie, R. B., Shindell, D. T., Strode, S. A., Sudo, K., Szopa, S., and Zeng, G.: Pre-industrial to end 21st century projections of tropospheric ozone from the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP), Atmos Chem Phys, 13, 2063-2090, 2013.
Yu, L., Xiangze Jin, and Weller, R. A.: Multidecade Global Flux Datasets from the Objectively Analyzed Air-sea Fluxes (OAFlux) Project: Latent and Sensible Heat Fluxes, Ocean Evaporation, and Related Surface Meteorological Variables. (OA-2008-01), W. H. O. I. O. P. T. R. (Ed.), 2008.
Zhang, Y. C., Rossow, W. B., Lacis, A. A., Oinas, V., and Mishchenko, M. I.: Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: Refinements of the radiative transfer model and the input data, J Geophys Res-Atmos, 109, 2004.
Zhu, Z. C., Bi, J., Pan, Y. Z., Ganguly, S., Anav, A., Xu, L., Samanta, A., Piao, S. L., Nemani, R. R., and Myneni, R. B.: Global Data Sets of Vegetation Leaf Area Index (LAI)3g and Fraction of Photosynthetically Active Radiation (FPAR)3g Derived from Global Inventory Modeling and Mapping Studies (GIMMS) Normalized Difference Vegetation Index (NDVI3g) for the Period 1981 to 2011, Remote Sens-Basel, 5, 927-948, 2013.
Ziemke, J. R., Chandra, S., Labow, G. J., Bhartia, P. K., Froidevaux, L., and Witte, J. C.: A global climatology of tropospheric and stratospheric ozone derived from Aura OMI and MLS measurements, Atmos Chem Phys, 11, 9237-9251, 2011.

Form
6 Pt.,

Table 1. Overview of standard namelists implemented in ESMValTool (v1.0) along with the quantity and ESMValTool variable name for which the namelist is tested, the corresponding observations or reanalyses, the section and example figure in this paper, and references for the namelist. When the namelist is named with a specific paper (naming convention: *namelist_SurnameYearJournalabbreviation.xml*), it can be used to reproduce in general all or in some cases only a subset of the figures published in that paper. Otherwise the namelists group a set of diagnostics and performance metrics for a specific scientific topic (e.g., *namelist_aerosol_CMIP5.xml*). Observations and reanalyses are listed together with their Tier, type (e.g., reanalysis, satellite or in situ observations), the time period used, and a reference. Tier 1 includes observations from obs4MIPs or reanalyses from ana4MIPs. Tier 2 and tier 3 indicate freely-available and restricted data sets, respectively. For these observations, reformatting routines are provided to bring the original data in the CF/CMOR standard format so that they can directly be used in the ESMValTool.

xml namelist	**Tested Quantity (CMOR units)**	**ESMValTool Variable Name**	**Tested Observations /Reanalyses (Tier, type, time period, reference)**	**Section / Example Figure(s)**	**References for namelist**
Section 4.1: Detection of systematic biases in the physical climate: atmosphere					
namelist_perfmetrics_CMIP5 *namelist_righi15gmd_ECVs*	Temperature (~~°C~~(K)) Eastward wind (m s^{-1}) Northward wind (m s^{-1}) Near-surface air temperature (~~°C~~(K)) Geopotential height (m)	ta ua va tas zg	ERA-Interim (Tier 3, reanalysis, 1979-2014 (Dee et al., 2011)) NCEP (Tier 2, reanalysis, 1948-2012 (Kistler et al., 2001))	Section 4.1.1. / Fig. 2 and Fig. 3	(Gleckler et al. (2008)); (Taylor (2001)); Fig. 9.7 of (Flato et al. (2013)) Righi et al. (2015)
	Specific Humidity (~~g kg~~1)	hus	AIRS (Tier 1, satellite, 2003-2010 (Aumann et al., 2003))		
	Precipitation (kg m^{-2} s^{-1})	pr	GPCP-SG (Tier 1, satellite & rain gauge, 1979-near-present (Adler et al., 2003))		
	TOA outgoing shortwave radiation (W m^{-2}) TOA outgoing longwave radiation (W m^{-2}) TOA outoing clear-sky longwave radiation (W m^{-2})	rsut rlut rlutcs	CERES-EBAF (Tier 1, satellite, 2001-2011 (Wielicki et al., 1996))		

	Shortwave cloud radiative effect (W m^{-2}) Longwave cloud radiative effect (W m^{-2})	SW_CRE LW_CRE			
	Aerosol optical depth at 550 nm (1)	od550aer	MODIS (Tier 1, satellite, 2001-2012 (King et al., 2003)) ESACCI-AEROSOL (Tier 2, satellite, 1996-2012 (Kinne et al., 2015))		
	Total cloud amount (%)	clt	MODIS (Tier 1, satellite, 2001-2012 (King et al., 2003))		
namelist_flato13ipcc	Near-surface air temperature (~~°C~~(K))	tas	ERA-Interim (Tier 3, reanalysis, 1979-2014 (Dee et al., 2011))	Section 4.1.2 / Fig. 4	Fig. 9.2 and Fig. 9.4 of (Flato et al. (2013))
	Precipitation (kg m^{-2} s^{-1})	pr	GPCP-1DD (Tier 1, satellite, 1997-2010 (Huffman et al., 2001))		
namelist_SAMonsoon *namelist_SAMonsoon_AMIP* *namelist_SAMonsoon_daily*	Eastward wind (m s^{-1}) Northward wind (m s^{-1})	ua va	ERA-Interim (Tier 3, reanalysis, 1979-2014 (Dee et al., 2011)) MERRA (Tier 1, reanalysis, 1979-2011 (Rienecker et al., 2011))	Section 4.1.3.1 / Fig. 5 and Fig. 6	(Goswami et al. (1999)) Sperber et al. (2013) ~~Wang and Fan (1999) Wang et al. (2012) Webster and Yang (1992)~~ Wang and Fan (1999) Wang et al. (2012) Webster and Yang (1992) Lin et al. (2008)~~;~~ ~~Fig. 9.32 of Flato et al. (2013);~~ Fig. 9.32 of Flato et al. (2013)
	Precipitation (kg m^{-2} s^{-1})	pr	TRMM-3B42-v7 (Tier 1, satellite, 1998-near-present (Huffman et al., 2007)) GPCP-1DD 1DD (Tier 1, satellite, 1997-2010 (Huffman et al., 2001)) CMAP (Tier 2, satellite & rain gauge, 1979-near-present (Xie and Arkin, 1997)) MERRA (Tier 1,		

			reanalysis, 1979-2011 (Rienecker et al., 2011)) ERA-Interim (Tier 3, reanalysis, 1979-2014 (Dee et al., 2011))		
	Skin temperature (K)	ts	HadISST (Tier 2, reanalysis, 1870-2014 (Rayner et al., 2003))		
namelist_WAMonsoon *namelist_WAMonsoon_daily*	Eastward wind (m s^{-1}) Northward wind (m s^{-1}) Temperature (~~°C~~(K) Near-surface air temperature (~~°C~~(K)	ua va ta tas	ERA-Interim (Tier 3, reanalysis, 1979-2014 (Dee et al., 2011))	Section 4.1.3.2 / Fig. 7	Roehrig et al. (2013)~~;~~ ~~Cook and Vizy (2006);~~ Cook and Vizy (2006)
	Precipitation (kg m^{-2} s^{-1})	pr	GPCP-1DD (Tier 1, satellite, 1997-2010 (Huffman et al., 2001)) TRMM (Tier 1, satellite, 1998-near-present (Huffman et al., 2007))		
	TOA outgoing shortwave radiation (W m^{-2}) TOA outgoing longwave radiation (W m^{-2}) TOA outoing clear-sky shortwave radiation (W m^{-2}) TOA outoing clear-sky longwave radiation (W m^{-2}) Shortwave cloud radiative effect (W m^{-2}) Longwave cloud radiative effect (W m^{-2}) Shortwave downwelling radiation at surface (W m^{-2}) Longwave downwelling radiation at surface (W m^{-2})	rsut rlut rsutcs rlutcs SW_CRE LW_CRE rsds rlds	CERES-EBAF (Tier 1, satellite, 2001-2011 (Wielicki et al., 1996))		
	TOA outgoing longwave radiation (W m^{-2})	rlut	NOAA polar-orbiting satellites (Tier 2, satellite,		

			1974- 2013 (Liebmann and Smith, 1996))		
namelist_CVDP	Precipitation (kg m^{-2} s^{-1})	pr	GPCP-SG (Tier 1, satellite & rain gauge, 1979-near-present (Adler et al., 2003)) TRMM (Tier 1, satellite, 1998-near-present (Huffman et al., 2007))	Section 4.1.4.1 / Fig. 8 and Fig. 9	(Phillips et al. (2014))
	Air pressure at sea level (Pa)	psl	NOAA-CIRES Twentieth Century Reanalysis Project (Tier 1, reanalysis, 1900-2012 (Compo et al., 2011))		
	Near-surface air temperature ~~(°C~~(K)	tas	NCEP (Tier 2, reanalysis, 1948-2012 (Kistler et al., 2001))		
	Skin temperature (K)	ts	HadISST (Tier 2, satellite-based, 1870-2014 (Rayner et al., 2003))		
	Snow depth (m)	snd	without obs		
	Ocean meridional overturning mass streamfunction (kg s^{-1})	msftmyz	without obs		
namelist_mjo_daily *namelist_mjo_mean_state*	Eastward wind (m s^{-1}) Northward wind (m s^{-1})	ua va	ERA-Interim (Tier 3, reanalysis, 1979-2014 (Dee et al., 2011)) NCEP (Tier 2, reanalysis, 1979-2013 (Kistler et al., 2001))	Section 4.1.4.2 / Fig. 10	~~Waliser et al. (2009);~~ (Kim et al. (2009); Waliser et al. (2009)); Kim et al. (2009)
	Precipitation (kg m^{-2} s^{-1})	pr	GPCP-1DD (Tier 1, satellite, 1997-2010 (Huffman et al., 2001))		
	TOA longwave radiation (W m^{-2})	rlut	NOAA polar-orbiting satellites (Tier 2, satellite, 1974- 2013 (Liebmann and Smith, 1996))		
namelist_~~diurnalcycle~~DiurnalCycle	Precipitation (kg m^{-2} s^{-1}) Convective Precipitation (kg m^{-2} s^{-1})	pr prc	TRMM (Tier 1, satellite, 1998-near-present (Huffman et al.,	Section 4.1.5 / Fig. 11	Rio et al. (2009)

			2007))		
	TOA outgoing longwave radiation (W m^{-2}) TOA outgoing shortwave radiation (W m^{-2}) TOA outgoing clear sky longwave radiation (~~clear sky) (~~W m^{-2}) TOA outgoing clear sky shortwave radiation (~~clear sky) (~~W m^{-2}) Surface downwelling shortwave radiation (W m^{-2}) Surface downwelling clear sky sky shortwave radiation (~~clear sky) (~~W m^{-2}) Surface upwelling shortwave radiation (W m^{-2}) Surface upwelling clear sky shortwave radiation (~~clear sky) (~~W m^{-2}) Surface upwelling longwave radiation (W m^{-2}) Surface upwelling clear sky longwave radiation (~~clear sky) (~~W m^{-2}) Surface downwelling shortwave radiation (W m^{-2}) Surface downwelling clear-sky longwave radiation (W m^{-2})	rlut rsut rlutcs rsutcs rsds rsdscs rsus rsuscs rlus rluscs rlds rldscs	CERES-SYN1deg (Tier 1, satellite, 2001-2011 (Wielicki et al., 1996))		
namelist_lauer13jclim	Atmosphere cloud condensed water content (kg m^{-2})	clwvi	UWisc: SSM/I, TMI, AMSR-E (Tier 3, satellite, 1988-2007 (O'Dell et al., 2008))	Section 4.1.6.1 / Fig. 12	Lauer and Hamilton (2013); Fig. ~~9.5 of Flato et al. (2013)~~9.5 of Flato et al. (2013)
	Atmosphere cloud ice content (kg m^{-2})	clivi	MODIS-CFMIP (Tier 2, satellite, 2003-2014 (King et al., 2003; Pincus et al., 2012))		
	Total cloud amount (%)	clt	MODIS (Tier 1, satellite, 2001-2012 (King et al., 2003))		

	TOA outgoing longwave radiation (W m^{-2}) TOA outgoing longwave radiation (clear sky) (W m^{-2}) TOA outgoing shortwave radiation (W m^{-2}) TOA outgoing shortwave radiation (clear sky) (W m^{-2})	rlut rlutcs rsut rsutcs	CERES-EBAF (Tier 1, satellite, 2001-2011 (Wielicki et al., 1996)) SRB (Tier 2, satellite, 1984-2007 (GEWEX-news, February 2011))		
	Precipitation (kg m^{-2} s^{-1})	pr	GPCP-SG (Tier 1, satellite & rain gauge, 1979-near-present (Adler et al., 2003))		
namelist_williams09climdyn_CREM	ISCPP mean cloud albedo (1) ISCCP mean cloud top pressure (Pa) ISCCP total cloud fraction (%) TOA outgoing shortwave radiation (W m^{-2}) TOA outgoing longwave radiation (W m^{-2}) TOA outoing clear-sky shortwave radiation (W m^{-2}) TOA outoing clear-sky longwave radiation (W m^{-2}) Surface snow area fraction (%) Surface snow amount (kg m^{-2}) Sea ice area fraction (%)	albisccp pctisccp cltisccp rsut rlut rsutcs rlutcs snc snw sic	ISCCP (Tier 1, satellite, 1985-1990 (Rossow and Schiffer, 1991)) ISCCP-FD (Tier 2, satellite, 1985-1990 (Zhang et al., 2004))	Section 4.1.6.2 / Fig. 13	Williams and Webb (2009)
Section 4.2: Detection of systematic biases in the physical climate: ocean					
namelist_SouthernOcean	Ocean Mixed Layer Thickness Defined by Sigma T (m)	mlotst	ARGO (Tier 2, Buoy, Monthly mean climatology 2001-2006 (Dong et al., 2008))	Section 4.2.2.1 / Fig. 14	CDFTOOLS
	Sea surface temperature (K) Downward heat flux at sea water surface (W m^{-2}) Surface Downward Eastward Wind Stress (Pa) Surface Downward Nordward Wind Stress (Pa)	~~Tos~~ tos hfds (hfls + hfss + rsns + rlns) tauu tauv	ERA-Interim (Tier 3, reanalysis, 1979-2014 (Dee et al., 2011))		

	Water Flux from precipitation and evaporation (kg m-2 s^{-1})	wfpe (pr + evspsbl)			
	Sea water salinity (psu) Sea surface salinity (psu) Sea Water Temperature (K)	so sos to	WOA09 (Tier 2, in-situ, climatology, (Antonov et al., 2010; Locarnini et al., 2010))		
	Sea Water X Velocity (m s^{-1}) Sea Water Y Velocity (m s^{-1})	uo vo	without obs		
namelist_SouthernHemisphere	Total ~~cloud amount~~Cloud Fraction (%) Atmosphere cloud ice content (kg m^{-2}) Atmosphere cloud condensed water content (kg m^{-2})	clt clivi clwvi	CloudSat (Tier 1, satellite, 2000-2005 (Stephens et al., 2002))	Section 4.2.2.2 / Fig. 15	(Frolicher et al. (2015))
	Surface upward latent heat flux (W m^{-2}) Surface upward sensible heat flux (W m^{-2})	hfls hfss	WHOI-OAflux (Tier 2, satellite-based, 2000-2005 (Yu et al., 2008))		
	TOA outgoing longwave radiation (W m^{-2}) TOA outgoing clear-sky longwave radiation (W m^{-2}) TOA outgoing shortwave radiation (W m^{-2}) TOA outgoing clear sky shortwave radiation (~~clear sky)~~ ~~(~~W m^{-2}) Surface downwelling shortwave radiation (W m^{-2}) Surface downwelling clear-sky longwave radiation (W m^{-2}) Surface downwelling shortwave radiation (W m^{-2}) Surface downwelling clear sky shortwave radiation (~~clear sky)~~ ~~(~~W m^{-2})	rlut rlutcs rsut rsutcs rlds rldscs rsds rsdscs	CERES-EBAF (Tier 1, satellite, 2001-2011 (Wielicki et al., 1996)) SRB (Tier 2, satellite, 1984-2007 (GEWEX-news, February 2011))		
namelist_TropicalVariability	Precipitation (kg m^{-2} s^{-1})	pr	TRMM (Tier 1, satellite, 1998-near-present (Huffman et al., 2007)	Section 4.2.3 / Fig. 16	~~Choi et al. (2011);~~ Choi et al. (2011); Li and Xie (2014)
	Sea surface temperature (K)	ts	HadISST (Tier 2, satellite-based, 1870-2014		

			(Rayner et al., 2003))		
	Eastward wind (m s^{-1}) Northward wind (m s^{-1})	ua va	ERA-Interim (Tier 3, reanalysis, 1979-2014 (Dee et al., 2011))		
namelist_SeaIce	Sea ice area fraction (%)	sic	HadISST (Tier 2, satellite-based, 1870-2014 (Rayner et al., 2003)) NSIDC (Tier 2, satellite, 1978-2010 (Meier et al., 2013; Peng et al., 2013))	Section 4.2.4 / Fig. 17	~~Stroeve et al. (2007) Stroeve et al. (2012); Fig. 9.24 of~~ (Flato et al. (2013); Stroeve et al. (2007)) Stroeve et al. (2012); Fig. 9.24 of Flato et al. (2013)
Section 4.3: Detection of systematic biases in the physical climate: land					
namelist_~~Evapotransport~~Evapotranspiration	Surface upward latent heat flux (W m^{-2})	hfls	LandFlux-EVAL (Tier 3, ground, 1989-2004 (Mueller et al., 2013)) GPCC (Tier 2, Rain gauge analysis, 1901-2010 (Becker et al., 2013))	Section 4.3.1 / Fig. 18	~~Mueller and Seneviratne (2014);~~ (Mueller and Seneviratne (2014); Orlowsky and Seneviratne (2013)); Orlowsky and Seneviratne (2013)
namelist_SPI	Precipitation (kg m^{-2} s^{-1})	pr	CRU (Tier 2, Rain gauge analysis, 1901-2010 (Mitchell and Jones, 2005))		
namelist_runoff_et	Total runoff (kg m^{-2} s^{-1}) Evaporation (kg m^{-2} s^{-1}) Precipitation (kg m^{-2} s^{-1})	mrro evspsbl pr	GRDC (Tier 2, river runoff gauges, varying periods (Dümenil Gates et al., 2000)) WFDEI (Tier 2, Reanalysis, 1979-2010 (Weedon et al., 2014))	Section 4.3.2 / Fig. 19	~~Dümenil Gates et al. (2000);~~ Dümenil Gates et al. (2000); Hagemann et al. (2013)~~;~~ ~~Weedon et al. (2014);~~ Weedon et al. (2014)
Section 4.4: Detection of biogeochemical biases: carbon cycle					
namelist_anav13jclim	Net biosphere production of carbon (kg m^{-2} s^{-1}) ~~Surface Downward CO_2 Flux into ocean (kg m^{-2} s^{-1})~~	nbp ~~fgco2~~	TRANSCOM (Tier 2, Reanalysis, 1985 - 2008 (Gurney et al., 2004))	Section 4.4.1 / Fig. 20 and Fig. 21	Anav et al. (2013)
	Gross primary production of	gpp	MTE (Tier 2,		

	carbon (mol m^{-2} s^{-1})		Reanalysis, 1982 - 2008 (Jung et al., 2009))		
	Leaf area index (mol m^{-2} s^{-1})	lai	LAI3g (Tier 2, Reanalysis, 1981 - 2008 (Zhu et al., 2013))		
	Carbon mass in vegetation (kg m^{-2})	cVeg	NDP-017b (Tier 2, remote sensing 2000 (Gibbs, 2006))		
	Carbon mass in soil pool (kg m^{-2})	cSoil	HWSD (Tier 2, reanalysis, climatology (Nachtergaele et al., 2012))		
	Primary organic Carbon Production by all types of phytoplankton (mol m^{-2} s^{-1})	intPP	SeaWiFS (Tier 2, satellite, 1998-2010 (Behrenfeld and Falkowski, 1997; McClain et al., 1998))		
	Near-surface air temperature (K)	tas	CRU (Tier 3, near-surface temperature analysis, 1901-2006)		
	Precipitation (kg m^{-2} s^{-1})	pr	CRU (Tier 2, rain gauge analysis, 1901-2010 (Mitchell and Jones, 2005))		
namelist_GlobalOcean	Surface partial pressure of CO_2 (~~μatm~~Pa)	spco2	SOCAT v2 (Tier 2, in-situ, 1968 - 2011 (Bakker et al., 2014)) ETH SOM-FFN (Tier 2, extrapolated in situ, 1998 - 2011, (Landschützer et al., 2014a, b))	Section 4.4.2 / Fig. ~~21~~22	
	Total chlorophyll mass concentration at surface (kg m^{-3})	chl	SeaWiFS (Tier 2, satellite, 1997 - 2010 (Behrenfeld and Falkowski, 1997; McClain et al., 1998))		
	Dissolved oxygen concentration (mol m^{-3})	o2	WOA05 (Tier 2, in situ, climatology 1950-2004 (Bianchi et al., 2012))		
	Total alkalinity at surface (mol m^{-3})	talk	T14 (Tier 2, in situ, 2005 (Takahashi et al.,		

			2014))		
Section 4.5: Detection of biogeochemical biases: chemistry and aerosols					
namelist_aerosol_CMIP5	Surface concentration of SO_4 (~~μg~~kg m^{-3}) Surface concentration of NO_3 (~~μg~~kg m^{-3}) Surface concentration of NH_4 (~~μg~~kg m^{-3}) Surface concentration of black carbon aerosol (~~μg~~kg m^{-3}) Surface concentration of dry aerosol ~~primary~~ organic matter (~~μg~~kg m^{-3}) Surface concentration of PM10 aerosol (kg m^{-3}) Surface concentration of PM2.5 aerosol (kg m^{-3})	~~concso4~~ ~~concno3~~ ~~concnh4~~ ~~concbc~~ ~~concoa~~ ~~concpm10~~ ~~concpm2p5~~sconcso4 sconcno3 sconcnh4 sconcbc sconcoa sconcpm10 sconcpm2p5	CASTNET (Tier 2, Ground, 1987 .2012 (Edgerton et al., 1990)) EANET (Tier 2, Ground, 2001-2005 (Totsuka et al., 2005)) EMEP (Tier 2, Ground, 1970-2014 IMPROVE (Tier 2, Ground, 1988-2014	Section 4.5.1 / Fig. ~~22~~23	Lauer et al. (2005) Aquila et al. (2011) Righi et al. (2013); Fig. ~~9.29 of Flato et al. (2013)~~9.29 of Flato et al. (2013)
	Aerosol Number Concentration (~~1~~m^{-3}) BC Mass Mixing Ratio (kg kg^{-1}) ~~Dry~~ Aerosol mass mixing ration (kg kg^{-1}) BC-Free Mass Mixing Ratio (kg kg^{-1})	~~conccnm~~ conccn mrbc mmraer mmrbcfree	Aircraft campaigns (Tier 3, aircraft, various)		
	Aerosol Optical Depth at 550 nm (1)	od550aer	AERONET (Tier 2, Ground, 1992-~~2012~~2015 (Holben et al.,		

			1998)) MODIS (Tier 1, satellite, 2001-2012 (King et al., 2003)) MISR (Tier ~~2~~1, Satellite, 2001-2012 (Stevens and Schwartz, 2012)) ESACCI-AEROSOL (Tier 2, satellite, ~~1996-2012~~1998-2011 (Kinne et al., 2015))		
namelist_righi15gmd_tropo3 *namelist_righi15gmd_Emmons*	Ozone (~~mol~~nmol mol^{-1})	tro3	Aura MLS-OMI (Tier 2, satellite, 2005-2013 (Ziemke et al., 2011)) Ozone sondes (Tier 2, sondes, 1995-2009 (Tilmes et al., 2011))	Section 4.5.2 / Fig. ~~23~~24	~~Ozone of Righi et al. (2015) including Emmons et al. (2000) diagnostic;~~ Emmons et al. (2000) Righi et al. (2015)
	Carbon Monoxide (mol mol^{-1})	vmrco	GLOBALVIEW (Tier 2, ground, 1991-2008, (GLOBALVIEW-CO2, 2008))		
	Nitrogen Dioxide (NOx = NO + NO2) (mol mol^{-1}) C2H4 Propane (mol mol^{-1}) C2H6 Propane (mol mol^{-1}) C3H6 Propane (mol mol^{-1}) C3H8 Propane (mol mol-1) CH3COCH3 Acetone (mol mol^{-1})	vmrnox vmrc2h4 vmrc2h6 vmrc3h6 vmrc3h8 vmrch3coch3	Emmons (Tier 2, aircraft, various campaign (Emmons et al., 2000))		
namelist_~~eyring06jgr~~eyring13jgr	Temperature (~~°C~~(K) Eastward wind (m s^{-1}) ~~Heat flux (v'T')~~	ta ua ~~vt100~~	ERA-Interim (Tier 3, reanalysis, 1979-2014 ~~(Dee et al., 2011)~~) NCEP (Tier 2, reanalysis, 1948-2012 (Kistler et al., 2001))	Section 4.5.2 / Fig. ~~24~~25	(Eyring et al. (2013); Eyring et al. (2006)): Fig. 9.10 of Flato et al. (2013)
	~~Temperature (°C)~~	~~ta~~	~~RATPAC (Tier 2,~~		

			~~radiosondes, Climatology (Free et al., 2005))~~ ~~ERA40 (Tier 3, reanalysis, 1979-1999 (Uppala et al., 2005))~~		
	~~Methane (mole mole^{-1})~~ ~~Hydrogen Chlorine (mole mole^{-1})~~ ~~Water vapour~~ ~~Age of Air (years)~~	~~ch4~~ ~~hcl~~ ~~h2o~~ ~~age~~	~~HALOE (Tier 2, satellite, 1991-2002 (Grooß and Russell Iii, 2005))~~		
	~~Inorganic Chlorine (mole mole^{-1})~~	~~cly~~	~~HCl estimates from Aura MLS (Tier 2, satellite, 2005-2013 (Ziemke et al., 2011)) and HALOE (Tier 2, campaign, 1991 – 2002 (Russell et al., 1993))~~		
	Total Column Ozone (DU)	toz	~~Ground-based (Tier 3, in-situ, climatology, (Fioletov et al., 2002))~~ ~~Merged satellite data (Tier 2, satellite, 1970-2014 (Stolarski and Frith, 2006))~~ ~~NIWA (Tier 3, sondes, climatology, (Bodeker et al., 2005))~~ ~~SBUV/2 (Tier 3, satellite, 1978-present (Miller et al., 2002))~~ ~~GOME/SCIA/GOME-2 (Tier 3, satellite, 1995-2013 (Loyola and Coldewey-Egbers, 2012))~~NIWA (Tier 3, sondes, climatology, Bodeker et al., 2005)		

	~~Ozone (mol mol^{-1})~~	~~tro3~~	~~AURA-MLS-OMI (Tier 2, satellite, 2005-2013 (Ziemke et al., 2011))~~		
~~*namelist_eyring13jgr*~~	~~Temperature (°C)~~ ~~Eastward wind (m s^{-1})~~	~~ta~~ ~~ua~~	~~ERA-Interim (Tier 3, reanalysis, 1979-2014 (Dee et al., 2011))~~ ~~NCEP (Tier 2, reanalysis, 1948-2012 (Kistler et al., 2001))~~	~~Section 4.6 / Fig. 25~~	~~Eyring et al. (2013): Fig. 9.10 of Flato et al. (2013)~~
	~~Temperature (°C)~~	~~ta~~	~~RATPAC (Tier 2, radiosondes, Climatology (Free et al., 2005))~~ ~~ERA40 (Tier 3, reanalysis, 1979-2014 (Uppala et al., 2005))~~		
	~~Total Column Ozone (DU)~~	~~toz~~	~~See above~~		
	Tropospheric column ozone (DU) Ozone (~~mole mole~~nmol mol^{-1})	tropoz tro3	AURA-MLS-OMI (Tier 2, satellite, 2005-2013 (Ziemke et al., 2011))		
Section 4.6: Linking model performance to projections					
namelist_wenzel14jgr	Near-surface air temperature ~~(°C~~(K)	tas	NCDC (Tier 2, reanalysis, 1880-2001 (Smith et al., 2008))	Section 4.~~7~~6 / Fig. 26	~~Wenzel et al. (2014); Fig. 9.45 of~~ Wenzel et al. (2014); Fig. 9.45 of (Flato et al. (2013))
	Net biosphere production of carbon (kg m^{-2} s^{-1}) Carbon Dioxide (mol mol^{-1}) Surface Downward CO_2 Flux into ocean (kg m^{-2} s^{-1})	nbp co2 fgco2	GCP (Tier 2, reanalysis, 1959-present, (Le Quéré et al., 2014))		

Table 2. Overview of the diagnostics included for each namelist along with specific calculations, the plot type, settings in the configuration file (cfg-file), and comments. See also Annex C in the Supplement for additional information.

xml namelist	Diagnostics included	Specific Calculations (e.g., statistical measures, regridding)	Plot Types	Settings in cfg-file	Comments
Section 4.1: Detection of systematic biases in the physical climate: atmosphere					
namelist_perfmetrics_CMIP5 *namelist_righi15gmd_ECVs*	perfmetrics_main.ncl	Time averages, Regional weighted averages, t-test for difference plots	Annual cycle line plot, zonal mean plot, lat-lon map plot	Specific plot type, time averaging (e.g. annual, seasonal and monthly climatologies, annual and multi-year monthly means), region, target grid, pressure level, reference model, difference plot (True/False), statistical significance level of t-test for difference plot, multi model mean/median	The results of the analysis are saved to a netCDF file for each model to be read by perfmetrics_grading.ncl or perfmetrics_taylor.ncl.
	perfmetrics_grading.ncl	Grading metric, normalization	No plot	~~Type~~Time averaging, region, pressure level, reference model, type of metric for grading models (RMSE, Bias) ~~Type~~type of normalization (mean, median, centered median)	For tractability the filename for every diagnostic is written into a temporary file, which then is read by the perfmetrics _XXX_collect.ncl scripts. Additional metric and normalization methods can be added.
	perfmetrics_taylor.ncl	~~Normalization~~Taylor metrics	No plot	~~Same as for perfmetrics_grading.ncl~~ Time averaging, region, pressure level, reference model	
	perfmetrics_grading_collect.ncl	Collection of model grades from pre-calculated netCDF files	Portrait diagram		If individual models did not provide output for all variables or are compared to a different number of observations, the code will recognize this and

					return a blank array entry, ~~prudeing~~producing a white box in the portrait diagram; produces Figure 9.7 included in *namelist_flato13ipcc*
	perfmetrics_taylor_collect.ncl	Collection of model grades from precalculated netCDF files	Taylor diagram		
namelist_flato13ipcc	clouds_ipcc.ncl	Multi-model means, linear regridding to the grid of the reference data set	Zonal mean plots, global map	Map projection (CylindricalEquidistant, Mercator, Mollweide), selection of target grid, time mean (annualclim, seasonal-clim), reference data set	~~Produces Figure 9.5 of Flato et al. (2013) with *namelist_flato13ipcc.nml*~~Produces Figure 9.5 of Flato et al. (2013) with *namelist_flato13ipcc*
	clouds_bias.ncl	Multi-model means, linear regridding to the grid of the reference data set	Global map	map projection (CylindricalEquidistant, Mercator, Mollweide), selection of target grid, time mean (annualclim, seasonal-clim), reference data set	~~Produces Figures 9.2 and 9.4 of Flato et al. (2013) with *namelist_flato13ipcc.xml*~~Produces Figures 9.2 and 9.4 of Flato et al. (2013) with *namelist_flato13ipccl*
namelist_SAMonsoon	SAMonsoon_wind_basic.ncl	Mean and interannual standard deviation	Map contour plot, regional mean, RMSE and spatial correlation are given in plot titles	Region (latitude, longitude), season (consecutive month), contour levels	Zonal and meridional wind fields are used; mean and standard deviation (across all years) for each model. This diagnostic also plots the difference of the mean/standard deviation with respect to a reference data set. Mean contour plots include wind vectors.
	SAMonsoon_wind_seasonal.ncl	Climatology, seasonal anomalies and interannual variability	Annual cycle	Region (latitude, longitude), season (consecutive month), line colours, multi model mean (y/n)	Dynamical indices calculated from zonal and meridional wind fields are used. Wind levels are selected by input quantity (e.g. ua-200-850 and va-200-850)
	SAMonsoon_precip_basic.ncl	Mean and interannual standard deviation	Map contour plot, regional mean, RMSE and spatial correlation are given in plot titles	Region (latitude, longitude), season (consecutive month), contour levels	Similar to SAMonsoon_wind_basic.ncl

	SAMonsoon_precip_seasonal.ncl	Climatology, seasonal anomalies and interannual variability	Annual cycle	Region (latitude, longitude), season (consecutive month), line colours, multi model mean (y/n)	Similar to SAMonsoon_wind_seasonal.ncl
	SAMonsoon_precip_domain.ncl	Mean and standard deviation	Map contour plot	Region (latitude, longitude), season (consecutive month), contour levels	Domain and intensity defined using summer and winter precipitation defined appropriately for each hemisphere. Differences from reference data set also plotted. Produces Figure 9.32 included in *namelist_flato13ipcc*
	SAMonsoon_teleconnections.ncl	Correlation between interannual seasonal mean Nino3.4 SST timeseries (5S-5N, 190-240E) and precipitation over monsoon region.	Map contour plot, regional mean, RMSE and spatial correlation are given in plot titles	Region (latitude, longitude), season (consecutive month), contour levels	pr and ts are used to calculate teleconnections between precip and interannual Nino3.4 SSTs. Differences from reference data set also plotted.
namelist_SAMonsoon_AMIP	SAMonsoon_wind_IAV.ncl	Mean and standard deviation	Time-series line plot	Region (latitude, longitude), season (consecutive month), multi model mean (y/n)	Seasonal means of dynamical indices calculated for each year from zonal and meridional wind fields are used.
	SAMonsoon_precip_IAV.ncl	Mean and standard deviation	Time-series line plot	Region (latitude, longitude), season (consecutive month), multi model mean (y/n)	Seasonal means of precipitation for each year are used. Note that the scripts in *namelist_SAMonsoon* and *namelist_SAMonsoon_daily* can be used for coupled and atmosphere-only models alike, but this namelist allows year-to-year variations to be examined only for atmosphere-only simulations forced by observed SSTs.
namelist_SAMonsoon_daily	SAMonsoon_precip_daily.ncl	Standard deviation of filtered daily precipitation rates for each season	Map contour plot. Regional mean, spatial correlation and averages for Bay of	Region (latitude, longitude), season (consecutive month), contour levels	Both, actual standard deviations and standard deviations normalized by a climatology (with masking for precipitation rates <

			Bengal (10-20N, 80-100E) and E. Eq. Indian Ocean (10S-10N, 80-10^{0}E) are given in plot titles.		1mm/day) are plotted.
	SAMonsoon_precip_propagation.ncl	Regional averages, lagged correlations, band-pass filtering of daily precipitation rates	Hovmöller diagrams: (lag, lat) and (lag, lon)	Regions (latitude, longitude), season (consecutive months), filter settings	Similar to *namelist_mjo_daily_propagation* but using 30-80 day band-pass filtering and regions appropriate for SASM.
namelist_WAMonsoon namelist_WAMonsoon_daily	WAMonsoon_contour_basic.ncl	Mean and standard deviation	Map contour plot	Region (latitude, longitude), season (consecutive months), specific contour levels	Similar to SAMonsoon_wind_basic.ncl
	WAMonsoon_wind_basic.ncl	Mean and standard deviation	Map contour and vector plot	Region (latitude, longitude), season (consecutive months), contour levels, reference vector length	Mean wind contour and vector plots at selected pressure level. Similar to SAMonsoon_wind_basic.ncl
	WAMonsoon_10W10E_1D_basic.ncl	Zonal average over 10°W-10°E	Latitude line plot	Region (latitude), season (consecutive month)	Only 2 dimensional fields
	WAMonsoon_10W10E_3D_basic.ncl	Zonal average over 10°W-10°E	Vertical profile (latitude vs. level) contour plot	Region (latitude, pressure level), season (consecutive month), contour levels	Only 3 dimensional fields
	WAMonsoon_precip_IAV.ncl	Seasonal anomalies and interannual variability	Time-series line plot	Region (latitude, longitude)	Similar to SAMonsoon_wind_IAV.ncl
	WAMonsoon_precip_seasonal.ncl	Mean annual cycle	Time-series line plot	Region (latitude, longitude)	Similar to SAMonsoon_wind_seasonal.ncl
	WAMonsoon_autocorr.ncl	1-day autocorrelation of 1-90d (intraseasonal) anomalies	Map contour plot	Region (latitude, longitude), season (consecutive months), filtering properties, contour levels	
	WAMonsoon_isv_filtered.ncl	Intra-seasonal variance (time filtering)	Map contour plot	Region (latitude, longitude), season (consecutive months), filtering properties, contour levels	
namelist_CVDP	cvdp_atmos.ncl	Renaming climo files to CVDP naming convention,	No plot		Needed for the CVDP coupling to the ESMValTool.

		Generates CVDP namelist with all models			
	cvdp_ocean.ncl	Renaming climo files to CVDP naming convention	No plot		
	cvdp_obs.ncl	Generates CVDP name-list with all observations	No plot	Reference model(s) for each variable	Needed for the CVDP coupling to the ESMValTool.
	cvdp_driver.ncl	Calls the CVDP	No plot		Needed for the CVDP coupling to the ESMValTool. Flexible implementation for easy update-processes, Results of the analysis are saved in netCDF files for each model/observation
	amo.ncl	Area-weighted average, linear regression, spectral analysis, regridding for area-weighted pattern correlation and RMS difference	Lat-lon contour plots, time-series, spectral plots		Original CVDP diagnostic
	amoc.ncl	Mean, standard deviation, EOF, linear regression, lag correlations, spectral analysis	Pattern plots, spectral plots, time-series		Original CVDP diagnostic
	pdo.ncl	EOF, linear regression, spectral analysis	Lat-lon contour plots, time-series, spectral plots		Original CVDP diagnostic
	pr.mean_stddev.ncl	Global means, standard deviation	Lat-lon contour plots		Original CVDP diagnostic
	pr.trends_timeseries.ncl	Global trends	Lat-lon contour plots, time-series		Original CVDP diagnostic
	psl.mean_stddev.ncl	Global means, standard deviation	Lat-lon contour plots		Original CVDP diagnostic
	psl.modes_indices.ncl	EOF, linear regression,	Lat-lon contour plots, time series		Original CVDP diagnostic
	psl.trends.ncl	Global trends	Lat-lon contour plots		Original CVDP diagnostic
	snd.trends.ncl	Global trends	Lat-lon contour plots		Original CVDP diagnostic

	sst.indices.ncl	Area-weighted average, standard deviation, spectral analysis	Spatial composites, ~~hovmoller~~Hovmöller diagram, time-series, spectral plots		Original CVDP diagnostic
	sst.mean_stddev.~~ne~~ncl	Global means, standard deviation	Lat-lon contour plots		Original CVDP diagnostic
	sst.trends_timeseries.ncl	Global trends	Lat-lon contour plots, time-series		Original CVDP diagnostic
	tas.mean_stddev.ncl	Global means, standard deviation	Lat-lon contour plots		Original CVDP diagnostic
	tas.trends_timeseries.ncl	Global trends	Lat-lon contour plots, timeseries		Original CVDP diagnostic
	metrics.ncl	Collect all area-weighted pattern correlations and RMS differences created by the various scripts, calculates total score	txt-file		Original CVDP diagnostic
	webpage.ncl	Creates webpages to display CVDP results	.html files		Original CVDP diagnostic
namelist_mjo_daily	mjo_wave_freq.ncl	Meridional averaged over 10°S-10°N, wavenumber-frequency	Wavenumber-frequency contour plot	Season (summer, winter), daily max/min, region (latitude)	
	mjo_univariate_eof.ncl	Conventional (covariance) univariate EOF analysis	Lat-lon contour plot	Region (latitude, longitude), number and name of EOF modes, contour levels	EOF for 20-100 day band-pass filtered daily anomaly data
	mjo_precip_u850-200_propagation.ncl	Correlation, zonal average over 80°E-100°E, meridional average over 10°S-10°N, reference region over 75°E-100°E,10°S-5°N	Lag-longitude and lag-latitude diagram	Season(summer, winter, annual), region(latitude, longitude)	Lead/lag correlation of two variables with daily time resolution
	mjo_precip_uwnd_variance.ncl	Variance	Lat-lon contour plot	Season (summer, winter), region (latitude, longitude), contour levels	20-100 day bandpass filtered variance for two variables with daily time resolution

	mjo_olr_u850-200_cross_spectra.ncl	Coherence squared and phase lag	Wavenumber-frequency contour plot	Region (latitude), segments length and overlapped segments length, spectra type	Missing values are not allowed in the input data
	mjo_olr_u850_200_ceof.ncl	CEOF	Line plot	Region(latitude),number and names of CEOF modes, y-axis limit	the first two CEOF modes (PC1 and PC2) are retained for the MJO composite life cycle analysis
	mjo_olr_uv850_ceof_life_cycle.ncl	Calculate mean value for each phase category	Lat-lon contour plot	Season (summer, winter), region (latitude, longitude)	The appropriate MJO phase categories are derived from PC1 and PC2 of CEOF analysis
namelist_mjo_mean_state	mjo_precip_u850_basic.ncl	Season mean	Lat-lon contour plot	Season (summer, winter), region (latitude, longitude)	Based on monthly data
namelist_~~diurnalcycle~~DiurnalCycle		Mean diurnal cycle computation, regridding of observations and models over a specific grid and first harmonic analysis to derive amplitude and phase of maximum rainfall	Composites of diurnal cycles over specific regions and seasons, global maps of maximum precipitation phase and amplitude		A prerequisite to use this namelist is to check the time axis of high frequency data from models and observations to be sure of what is provided. One should check in particular if it is instantaneous or averaged values, and if the time provided corresponds to the middle or the end of the 3h interval. Note that timeaxis is modified in the namelist to make data coherent.
namelist_lauer13jclim	clouds.ncl	Multi-model mean	Lat-lon contour plot	map projection (CylindricalEquidistant, Mercator, Mollweide), destination grid	Produces Figure 9.5 included in *namelist_flato13ipcc*
	clouds_taylor.ncl	Multi-model mean	Taylor diagram		Taylor diagrams
	clouds_interannual.ncl	Interannual variability, multi-model mean	Lat-lon contour plot	Map projection (CylindricalEquidistant, Mercator, Mollweide), destination grid, reference data sets	
namelist_williams09climdyn_CREM	ww09_ESMValTool.py	Model data assigned to observed cloud regimes and regime frequency and mean radiative properties calculated.	Bar graph		

Section 4.2: Detection of systematic biases in the physical climate: ocean					
namelist_SouthernOcean	SeaIce_polcon.ncl		Polar stereographic maps	contour values	
	SeaIce_polcon_diff.ncl	Rregridding (ESMF)	Polar stereographic maps	contour values, reference model	
	SouthernOcean_vector_polcon_diff.ncl	Vector overlay (magnitude and direction)	Polar stereographic maps	contour plot scales, reference model	based on SeaIce_polcon_diff.ncl, variables with u and v components
	SouthernOcean_areamean_vertconplot.ncl	Regridding (ESMF)	Zonal mean vertical profiles (~~Hovmoller~~Hovmöller diagrams)	coordinates of subdomain	based on CDFTOOLS package
	SouthernOcean_transport.ncl	Sea water volume transport calculation	Line plot	coordinates of subdomain	
namelist_SouthernHemisphere	SouthernHemisphere.py	Regridding (interpolation to common grid), Temporal and zonal averages, RMSEs	Seasonal cycle line plot with calculated RMSEs and zonal mean contour plot	Masking of unwanted values (limits), region (coordinates) and season (months) specification, plotting limits, contour colourmap	
	SouthernHemisphere_scatter.py	Covariability of radiation fluxes as function of cloud metrics	Scatter plot of values with line plot of value distribution		
namelist_TropicalVariability	TropicalVariability.py	Temporal and zonal averages, RMSEs, normalization, co-variability	Annual cycles, seasonal scatter plots with calculated RMSEs	Masking of unwanted values (limits), Region (coordinates) and season (months), plotting limits	Fig. 5 of Lie and Xie, 2014
	TropicalVariability_EQ.py	Temporal and zonal averages, RMSEs, normalization, co-variability	Latitude cross sections of equatorial variables		
	TropicalVariability_wind.py	Regridding (interpolation)	Wind divergence plots		
namelist_SeaIce	SeaIce_tsline.ncl	Sea-ice area and extent, regridding (ESMF)	Time series	~~selection~~Selection of Arctic/Antarctic,	Produces Figure 9.24 included in *namelist_flato13ipcc*
	SeaIce_ancyc.ncl	Sea-ice area and extent, regridding (ESMF)	Annual cycle line plot	~~selection~~Selection of Arctic/Antarctic	
	SeaIce_polcon.ncl	Sea-ice area and extent, regridding	Polar stereographic maps	~~selection~~Selection of Arctic/Antarctic, optional red line	

		(ESMF)		depicting edges of sea-ice extent	
	SeaIce_polcon_diff.ncl	Sea-ice area and extent, regridding (ESMF)	Polar stereographic maps	~~selection~~Selection of Arctic/Antarctic, optional red line depicting edges of sea-ice extent	
Section 4.3: Detection of systematic biases in the physical climate: land					
namelist_~~Evapotransport~~Evapotranspiration	Evapotranspiration.ncl	Conversion to evapotranspiration units, global average, RMSE	Lat-lon contour plot	Time period	
namelist_SPI	SPI.r	SPI calculation	Lat-lon contour plot	Time period, time scale (3, 6 or 12 monthly)	May require manual installation of certain R-packages to run
namelist_runoff_et	catchment_analysis_val.py	Temporal and spatial mean for 12 large river catchments, regridding to 0.5x0.5 lat-lon grid	Bar plots of evapotranspiration and runoff bias against observation, scatter plots of runoff bias against the biases of evapotranspiration precipitation	(no cfg. file)	Three variables are read by this diagnostic.
Section 4.4: Detection of biogeochemical biases: carbon cycle					
namelist_anav13jclim	Anav_MVI_IAV_Trend_Plot.ncl	Regridding to common grid, monthly and annual special averages , variability (MVI = (model/reference - reference/model) 2)	Scatter plot	Region (latitude), resolution size for regridding (e.g., 0.5°, 1°, 2°)	All carbon flux variables were corrected for the exact amount of carbon in the coastal regions by applying the models land-ocean fraction to the variables.
	Anav_Mean_IAV_ErrorBars_Seasonal_cycle_plots.ncl	Regridding to common grid Monthly and annual special averages	Seasonal cycle line plot, scatter plot, error-bar plot	Region (latitude), resolution size for regridding (e.g., 0.5°, 1°, 2°)	
	Anav_cSoil-cVeg_Scatter.ncl	Regridding to common grid annual special averages	Scatter plot	Region (latitude), resolution size for regridding (e.g., 0.5°, 1°, 2°)	Two variables are read by this diagnostic
	perfmetrics_grading.ncl	RMSE, PDF-skill score	No plot		See details in *namelist_perfmetrics_CMIP5*
	perfmetrics_grading_collect.ncl		Portrait diagram		See details in *namelist_perfmetrics_CMIP5*
namelist_GlobalOcean	GO_tsline.ncl	Multi-model mean	Time-series line plot	Region (lat/lon), pressure levels, optional smoothing,	

				anomaly calculations, overlaid trend lines, and masking of model data according to observations	
	GO_comp_map.ncl	Mean, standard deviation, and difference to reference model	Lat-lon contour plot (for specified *z*-level)	Region (Lat/lon), ocean depth, contour levels	Actual metrics ported from UK MetOffice IDL-monsoon evaluation scripts
Section 4.5: Detection of biogeochemical biases: chemistry and aerosols					
namelist_aerosol_CMIP5	aerosol_stations.ncl	Collocation of model and observational data~~Regridding to coarsest grid~~	Time series, scatter plot, map plot	~~Observed stationdata is specified in the cfg-file~~Time averaging, station data network	All available observational data in the selected time period, on a monthly-mean basis is considered. The model data is extracted in the grid boxes where the respective observational stations are located (~~co-located~~collocated model and observational data). ~~Reproducing Figure 9.29 also with *namelist_flato13ipcc*~~
	aerosol_satellite.ncl	Regridding to coarsest grid	Map plots and difference plots	Target grid	
	aerosol_profiles.ncl	Mean, standard deviation, median, 5-10-25-75-90-95 percentiles	Vertical profiles		The model data are extracted based on the campaign/station location (lat-lon box) and time period (on a climatological basis, i.e. selecting the same days/months, but regardless of the year). Rather specific variables are required (i.e., aerosol number concentration for particles with diameter larger than 14 nm) to match the properties of the instruments used during the campaign.
	tsline.ncl		Line plot	Time averaging (annual, seasonal and monthly climatologies, annual and multi-year monthly	

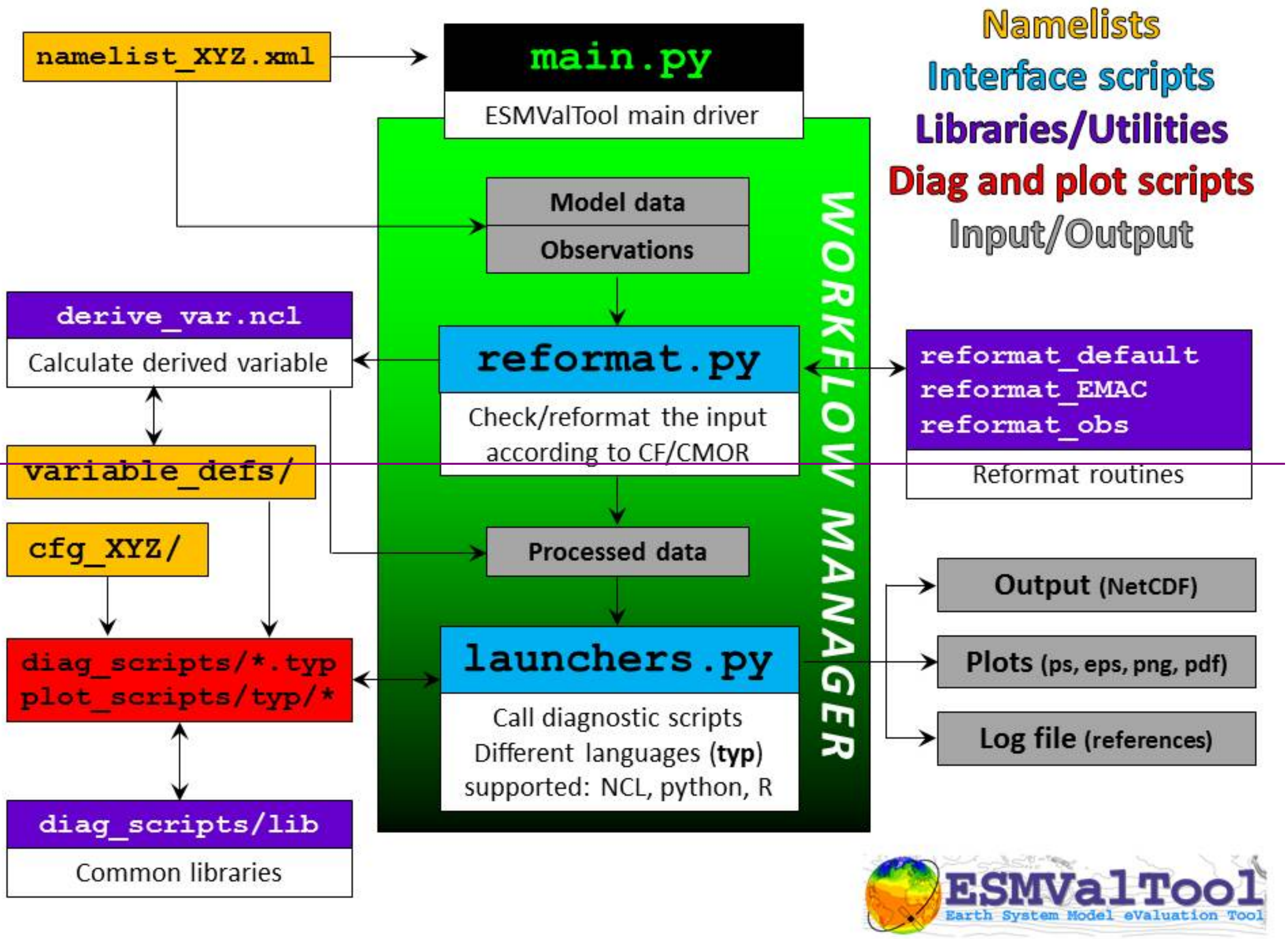
				means~~)~~ ~~Region~~), region (latitude, longitude)	
namelist_righi15gmd_tropo3	ancyc_lat.ncl	Regridding to reference global (area-weighted) average, zonal mean	Seasonal Hovmöller (month vs. latitude)		global (area-weighted) average is calculated only for grid cells with available observational data
	lat_long.ncl	Regridding to coarsest grid global (area-weighted) average			global (area-weighted) average is calculated only for grid cells with available observational data
	perfmetrics_main.ncl		Annual cycle line plot, zonal mean plot, lat-lon map plot		See details in *namelist_perfmetrics_CMIP5*
	perfmetrics_grading.ncl		No plot		See details in *namelist_perfmetrics_CMIP5*
	perfmetrics_taylor.ncl		No plot		See details in *namelist_perfmetrics_CMIP5*
	perfmetrics_grading_collect.ncl		Portrait diagram		See details in *namelist_perfmetrics_CMIP5*
	perfmetrics_taylor_collect.ncl		Taylor diagram		See details in *namelist_perfmetrics_CMIP5*
namelist_righi15gmd_Emmons	Emmons.ncl	Percentiles (5,25,75,95)%	Vertical profiles	~~Reference/~~Name(s) of the observational ~~profile file must be specified~~campaign(s)	
~~*namelist_eyring06jgr*~~	~~eyring06jgr_fig01.ncl~~	~~Climatological mean bias~~	~~Vertical profiles~~	~~Multi model mean (True/False), regions (latitude, longitude), time averaging (annual, individual month, seasons)~~	
	~~eyring06jgr_fig02.ncl~~	~~Cosine weighting for latitude averaging~~		~~Multi model mean (True/False) , regions (latitude, longitude), time averaging (annual, individual month, seasons)~~	
	~~eyring06jgr_fig03.ncl~~	~~Linear regression~~	~~Scatter plot and correlation coefficient~~	~~Multi model mean (True/False), tegions (latitude, longitude), pressure level, time averaging (annual, individual month, seasons)~~	~~Two variables are read.~~

	~~eyring06jgr_fig04.ncl~~	~~Anomalies with respect to first 10 years~~	~~Time seris~~	~~Multi model mean (True/False), anomaly (True/False), regions (latitude, longitude), time averaging (annual, individual month, seasons)~~	
	~~eyring06jgr_fig12b.ncl~~	~~Anomalies with respect to first 10 years~~	~~Time series~~	~~Multi model mean (True/False), regions (latitude, longitude), time averaging (annual, individual month, seasons)~~	
	~~eyring06jgr_fig05.ncl~~	~~Climatological mean~~	~~Zonal mean vertical profile~~	~~Multi model mean (True/False), regions (latitude, longitude), time averaging (annual, individual month, seasons)~~	
	~~eyring06jgr_fig07.ncl~~	~~Seasonal cycle averages~~	~~Seasonal cycle line plot~~	~~Multi model mean (True/False), regions (latitude, longitude), pressure level, time averaging (annual, individual month, seasons)~~	
	~~eyring06jgr_fig08.ncl~~	~~Cosine weighted area average, seasonal average~~	~~Seasonal Hovmöller (month vs. latitude)~~	~~Multi model mean (True/False), regions (latitude, longitude), time averaging (annual, individual month, seasons)~~	~~Similar to ancyc_lat.ncl: seasonal Hovmöller (month vs. latitude) diagrams are created but showing the moth for two years in a row for improved analysis of the periodicity.~~
	~~eyring06jgr_fig09.ncl~~	~~Phase lag and relative amplitude of annual cycles~~	~~Vertical profiles~~	~~Multi model mean (True/False), regions (latitude), time averaging (annual, individual month, seasons)~~	
	~~eyring06jgr_fig12.ncl~~	~~Cosine weighted area average, time average~~	~~Vertical profile, time series~~	~~Multi model mean (True/False), regions (latitude, longitude), time averaging (annual, individual month, seasons)~~	
	~~eyring06jgr_fig14.ncl~~	~~Zonal mean, seasonal average~~	~~Seasonal Hovmöller (month vs. latitude)~~	~~Multi model mean (True/False), regions (latitude, longitude), time averaging (annual,~~	~~Similar calculation as for ancyc_lat.ncl~~

				~~individual month, seasons)~~	
	~~eyring06jgr_fig15.ncl~~	~~Anomalies with respect to first 10 years, seasonal cycle mean~~	~~Time seris, seasonal cycle line plot~~	~~Multi model mean (True/False), regions (latitude, longitude), pressure level, time averaging (annual, individual month, seasons)~~	~~Similar to eyring06jgr_fig04.ncl anomalies of time series are generated but are compared to the seasonal cycle of the quantity in an extra panel.~~
namelist_eyring13jgr	ancyc_lat.ncl		Seasonal Hovmöller (month vs. latitude)		See details in *namelist_righi15gmd_tropo3*
	eyring13jgr_fig01.ncl		Seasonal Hovmöller (month vs. latitude)	Multi model mean (True/False), regions (latitude, longitude), time averaging (annual, individual month, seasons)	
	eyring13jgr_fig02.ncl		Time series	Multi model mean (True/False), regions (latitude, longitude), time averaging (annual, individual month, seasons)	Produces Figure 9.10 of Flato et al. (2013) included in *namelist_flato13ipcc*
	eyring13jgr_fig04.nxl	Tropospheric column ozone	Global maps		
	eyring13jgr_fig06.ncl	Anomalies with respect to a specifiable base line, mean and standard deviation (95% confidence) for simulation experiment	Time series	Multi model mean (True/False), regions (latitude, longitude), time averaging (annual, individual month, seasons)	
	eyring13jgr_fig07.ncl	Mean simulation experiments, differences between future scenario simulations and historical simulations	Vertical profile	Multi model mean (True/False), regions (latitude, longitude), time averaging (annual, individual month, seasons), list of models w/o interactive chemistry	
	eyring13jgr_fig10.ncl	Time averages, linear trends	Error bar plot	Multi model mean (True/False), regions (latitude, longitude), height (in km), time averaging (annual, individual month, seasons)	
	eyring13jgr_fig11	Correlations and	Scatterplot	Multi model mean	Two quantities are

	.ncl	correlation coefficient		(True/False), regions (latitude, longitude), time averaging (annual, individual month, seasons)	compared to each other for individual models and simulations at once. Simulations are indicated by different marker types.
Section 4.6: Linking model performance to projections					
namelist_wenzel14jgr	tsline.ncl	Cosine weighting for latitude averaging, anomaly with respect to first 10 years	Line plot	Multi model mean (True/False), anomaly (True/False), regions (latitude, longitude), time averaging (annual, individual month, seasons)	
	carbon_corr_2vars.ncl	Linear regression	Scatter plot and correlation coefficient	Exclude two years after volcanic eruptions (True/False: Mount Agung, 1963; El Chichon, 1982; and Mount Pinatubo, 1991)	Two variables are read. The gradient of the linear regression and the prediction error of the fit, giving γ_{IAV}, are saved in an external netCDF file to be read by the *carbon_constraint.ncl* script.
	carbon_constraint.ncl	$\gamma_{LT} = \frac{\Delta nbp^{c} - \Delta nbp^{u}}{\Delta tas^{c}}$ 'c'coupled simulation 'u' biocemically coupled simulation Gaussian-Normal PDF Conditional PDF	Scatter plot and correlation coefficient	Time period, region (latitude)	Three variables are read. (1) γ_{LT} is diagnosed from the models (2) the previously saved netCDF files containing γ_{IAV} values are read and correlated to γ_{LT} (3) normal and conditional PDFs for the pure model ensemble and the constraint γ_{LT} values are calculated Produces Figure 9.45 included in *namelist_flato13ipcc*

FIGURES


namelist_XYZ.xml
main.py
ESMValTool main driver
Namelists
Interface scripts
Libraries/Utilities
Diag and plot scripts
Input/Output
Model data
Observations
WORKFLOW MANAGER
derive_var.ncl
Calculate derived variable
reformat.py
Check/reformat the input according to CF/CMOR
reformat_default
reformat_EMAC
reformat_obs
Reformat routines
variable_defs/
cfg_XYZ/
Processed data
Output (NetCDF)
Plots (ps, eps, png, pdf)
Log file (references)
diag_scripts/*.typ
plot_scripts/typ/*
launchers.py
Call diagnostic scripts
Different languages (typ)
supported: NCL, python, R
diag_scripts/lib
Common libraries
ESMValTool
Earth System Model eValuation Tool

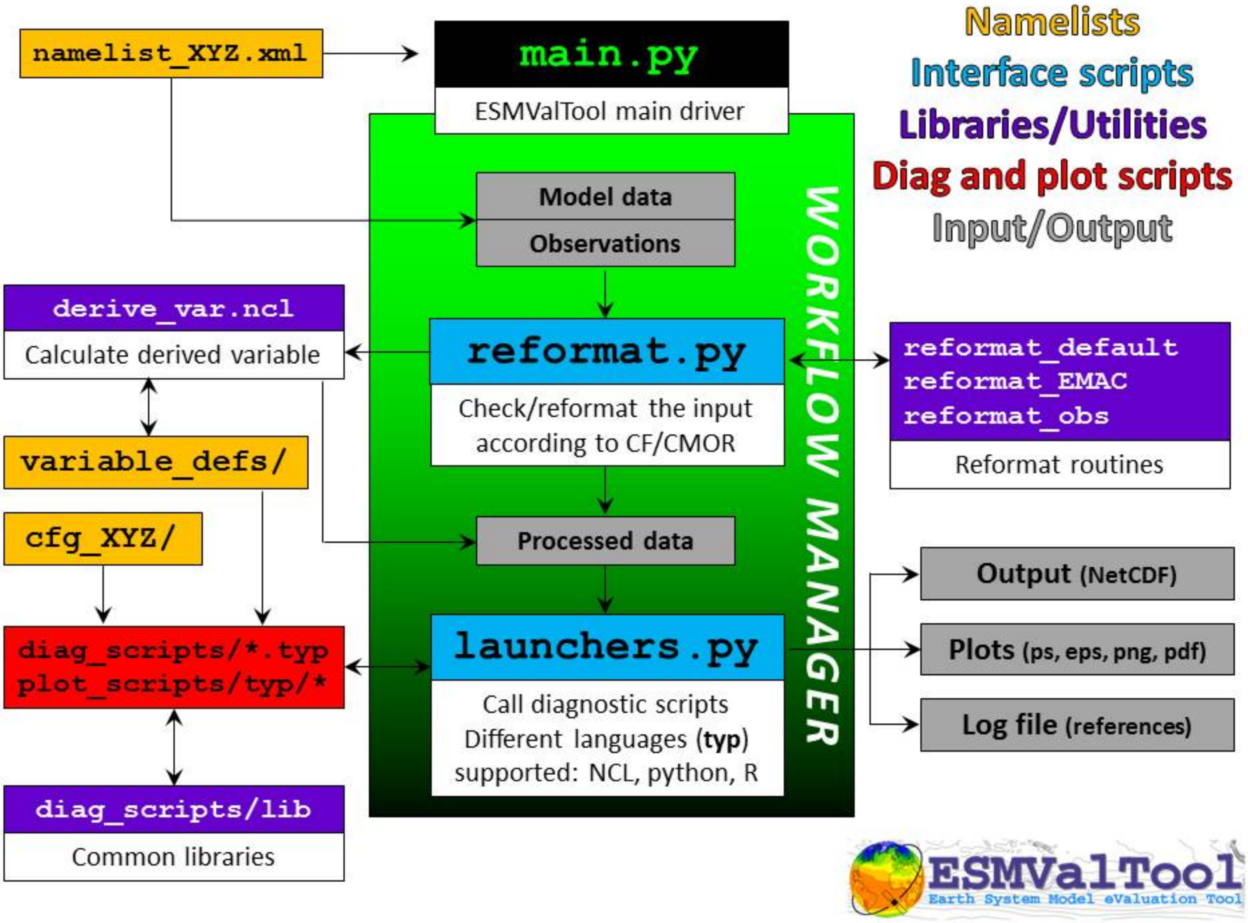
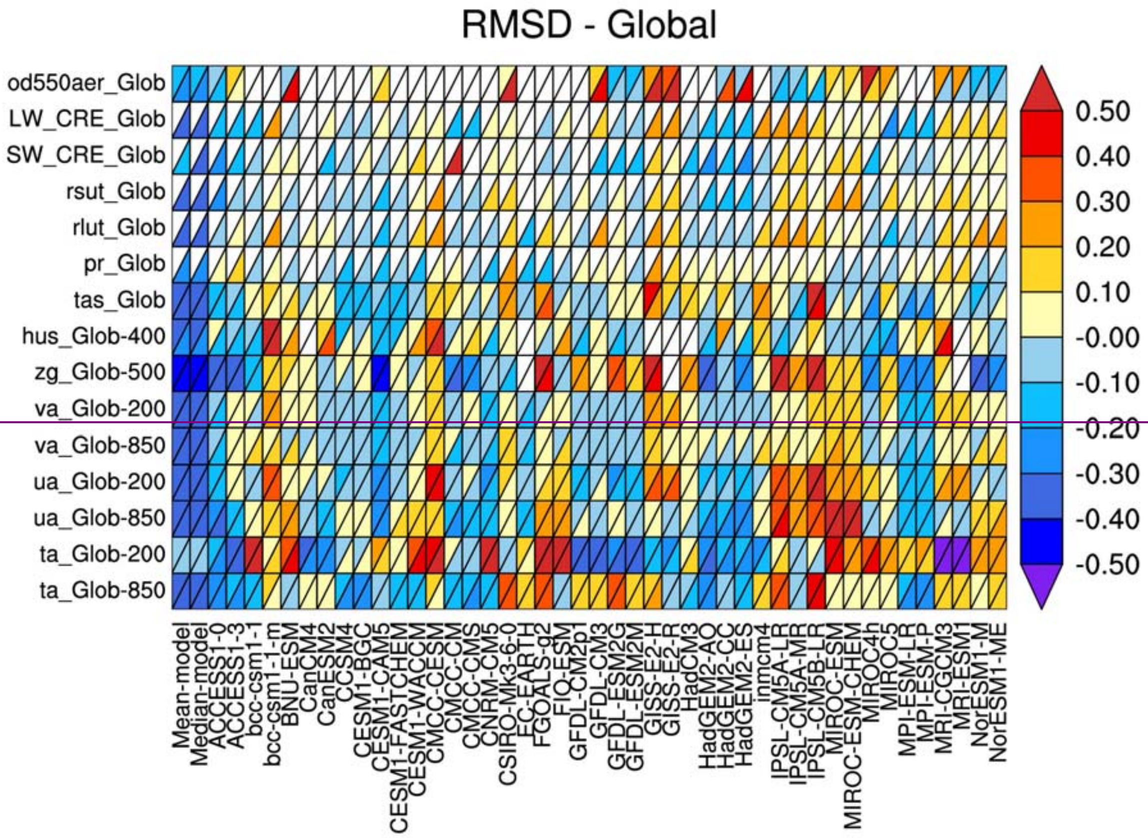


Figure 1. Schematic overview of the ESMValTool (v1.0) structure. The primary input to the workflow manager is a user-configurable text namelist file (orange). Standardized libraries/utilities (purple) available to all diagnostics scripts are handled through common interface scripts (blue). The workflow manager runs diagnostic scripts (red) that can be written in several freely-available scripting languages. The output of the ESMValTool (gray) includes figures, binary files (netCDF), and a log-file with a list of relevant references and processed input files for each diagnostic~~.~~”.


RMSD - Global
od550aer_Glob
LW_CRE_Glob
SW_CRE_Glob
rsut_Glob
rlut_Glob
pr_Glob
tas_Glob
hus_Glob-400
zg_Glob-500
va_Glob-200
va_Glob-850
ua_Glob-200
ua_Glob-850
ta_Glob-200
ta_Glob-850
0.50
0.40
0.30
0.20
0.10
-0.00
-0.10
-0.20
-0.30
-0.40
-0.50
Mean-model
Median-model
ACCESS1-0
ACCESS1-3
bcc-csm1-1
bcc-csm1-1-m
BNU-ESM
CanCM4
CanESM2
CCSM4
CESM1-BGC
CESM1-CAM5
CESM1-FASTCHEM
CESM1-WACCM
CMCC-CESM
CMCC-CM
CMCC-CMS
CNRM-CM5
CSIRO-Mk3-6-0
EC-EARTH
FGOALS-g2
FIO-ESM
GFDL-CM2p1
GFDL-CM3
GFDL-ESM2G
GFDL-ESM2M
GISS-E2-H
GISS-E2-R
HadCM3
HadGEM2-AO
HadGEM2-CC
HadGEM2-ES
inmcm4
IPSL-CM5A-LR
IPSL-CM5A-MR
IPSL-CM5B-LR
MIROC-ESM
MIROC-ESM-CHEM
MIROC4h
MIROC5
MPI-ESM-LR
MPI-ESM-P
MRI-CGCM3
MRI-ESM1
NorESM1-M
NorESM1-ME

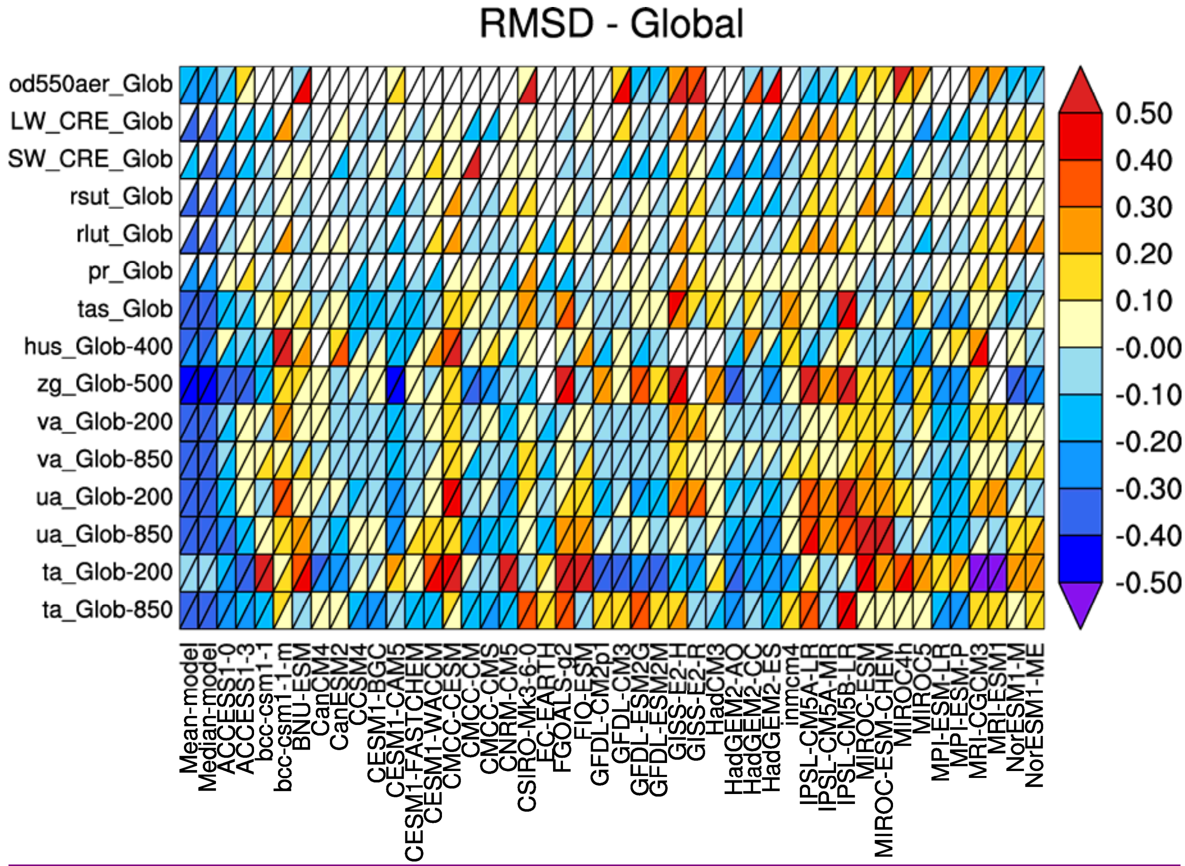
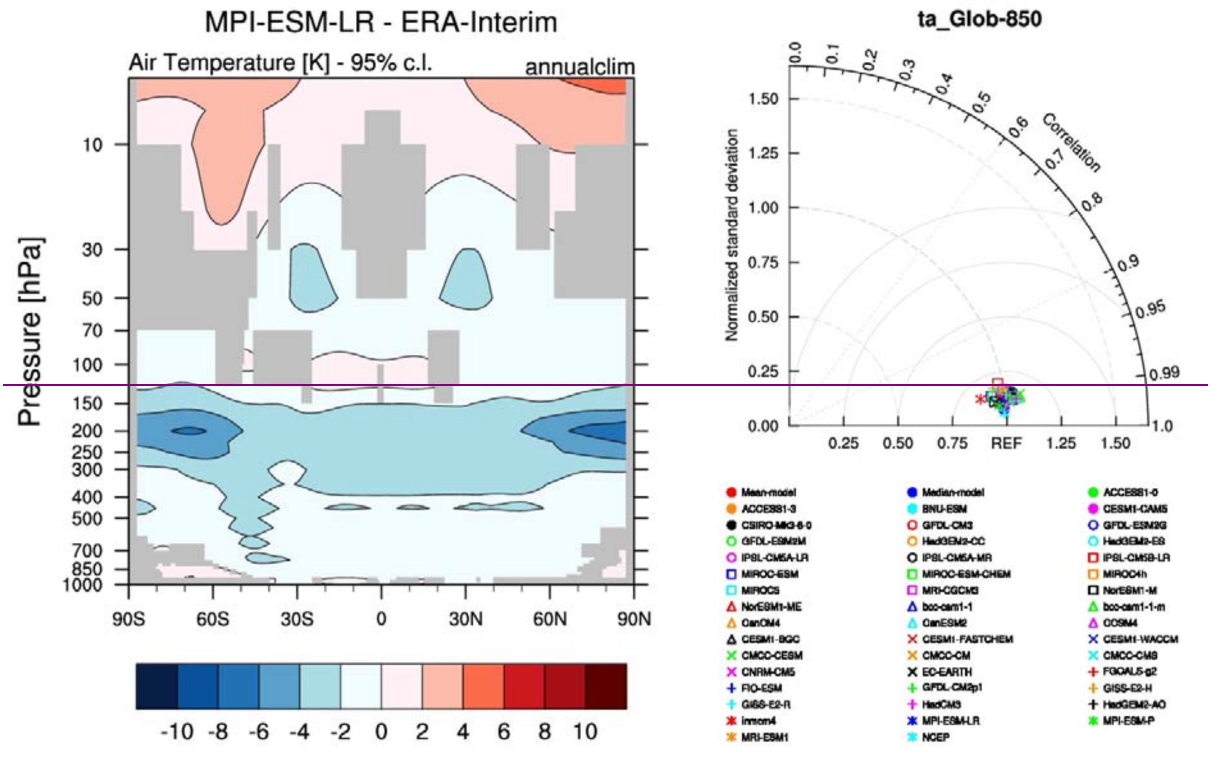

RMSD - Global
od550aer_Glob
LW_CRE_Glob
SW_CRE_Glob
rsut_Glob
rlut_Glob
pr_Glob
tas_Glob
hus_Glob-400
zg_Glob-500
va_Glob-200
va_Glob-850
ua_Glob-200
ua_Glob-850
ta_Glob-200
ta_Glob-850
0.50
0.40
0.30
0.20
0.10
-0.00
-0.10
-0.20
-0.30
-0.40
-0.50
Mean-model
Median-model
ACCESS1-0
ACCESS1-3
bcc-csm1-1
bcc-csm1-1-m
BNU-ESM
CanCM4
CanESM2
CCSM4
CESM1-BGC
CESM1-CAM5
CESM1-FASTCHEM
CESM1-WACCM
CMCC-CESM
CMCC-CM
CMCC-CMS
CNRM-CM5
CSIRO-Mk3-6-0
EC-EARTH
FGOALS-g2
FIO-ESM
GFDL-CM2p1
GFDL-CM3
GFDL-ESM2G
GFDL-ESM2M
GISS-E2-H
GISS-E2-R
HadCM3
HadGEM2-AO
HadGEM2-CC
HadGEM2-ES
inmcm4
IPSL-CM5A-LR
IPSL-CM5A-MR
IPSL-CM5B-LR
MIROC-ESM
MIROC-ESM-CHEM
MIROC4h
MIROC5
MPI-ESM-LR
MPI-ESM-P
MRI-CGCM3
MRI-ESM1
NorESM1-M
NorESM1-ME

Figure 2. Relative space-time root-mean square error (RMSE) calculated from the 1980–2005 climatological seasonal cycle of the CMIP5 historical simulations. A relative performance is displayed, with blue shading indicating performance being better and red shading worse, than the median of all model results. A diagonal split of a grid square shows the relative error with respect to the reference data set (lower right triangle) and the alternate data set (upper left triangle). White boxes are used when data ~~is~~are not available for the given model and variable or no alternate data set has been used. The figure shows that performance varies across CMIP5 models and variables, with some models comparing better with observations for one variable and another model performing better for a different variable. Except for global average temperatures at 200 hPa where most but not all models have a systematic bias, the multi-model mean outperforms any individual model. ~~Similar to Gleckler et al. (2008) and Figure 9.7 of Flato et al. (2013)~~Similar to Gleckler et al. (2008) and Figure 9.7 of Flato et al. (2013) produced with *namelist_perfmetrics_CMIP5.xml*.



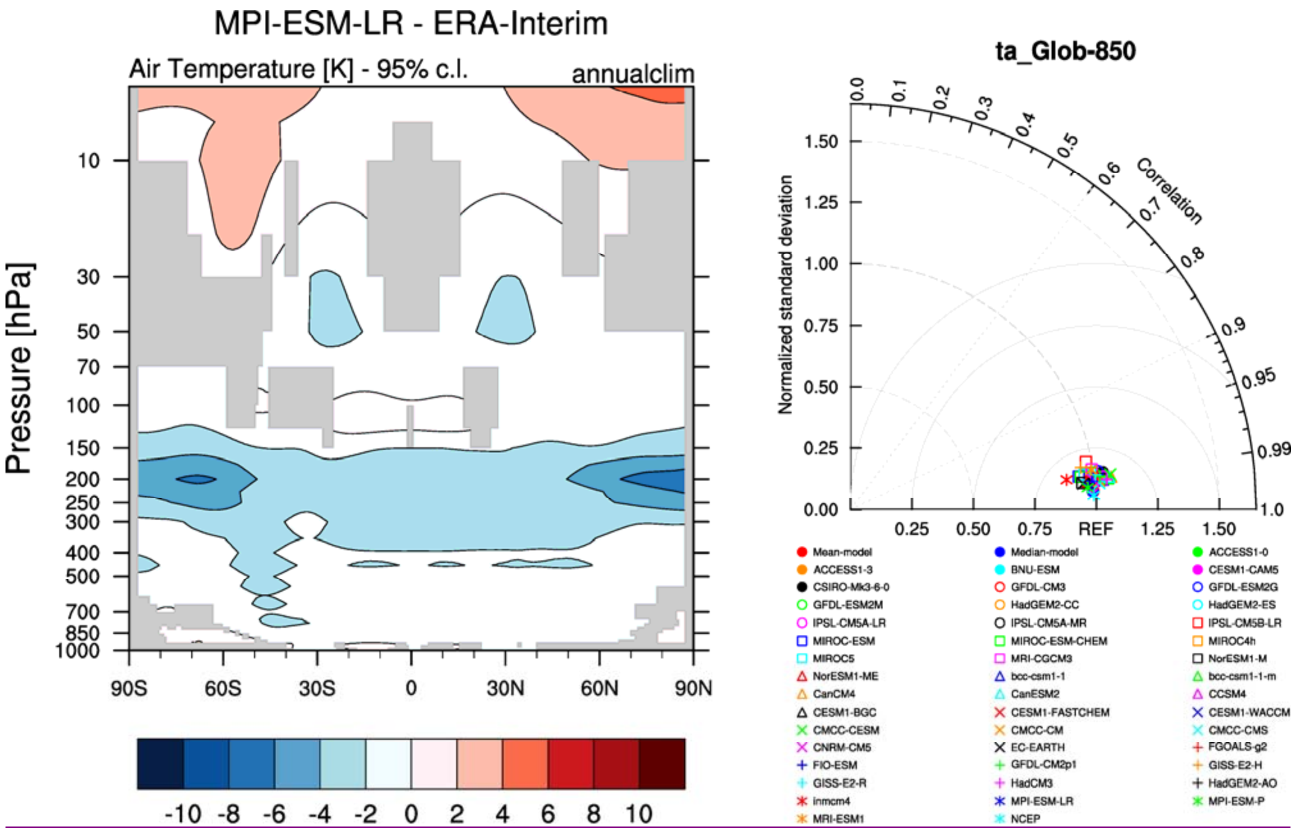
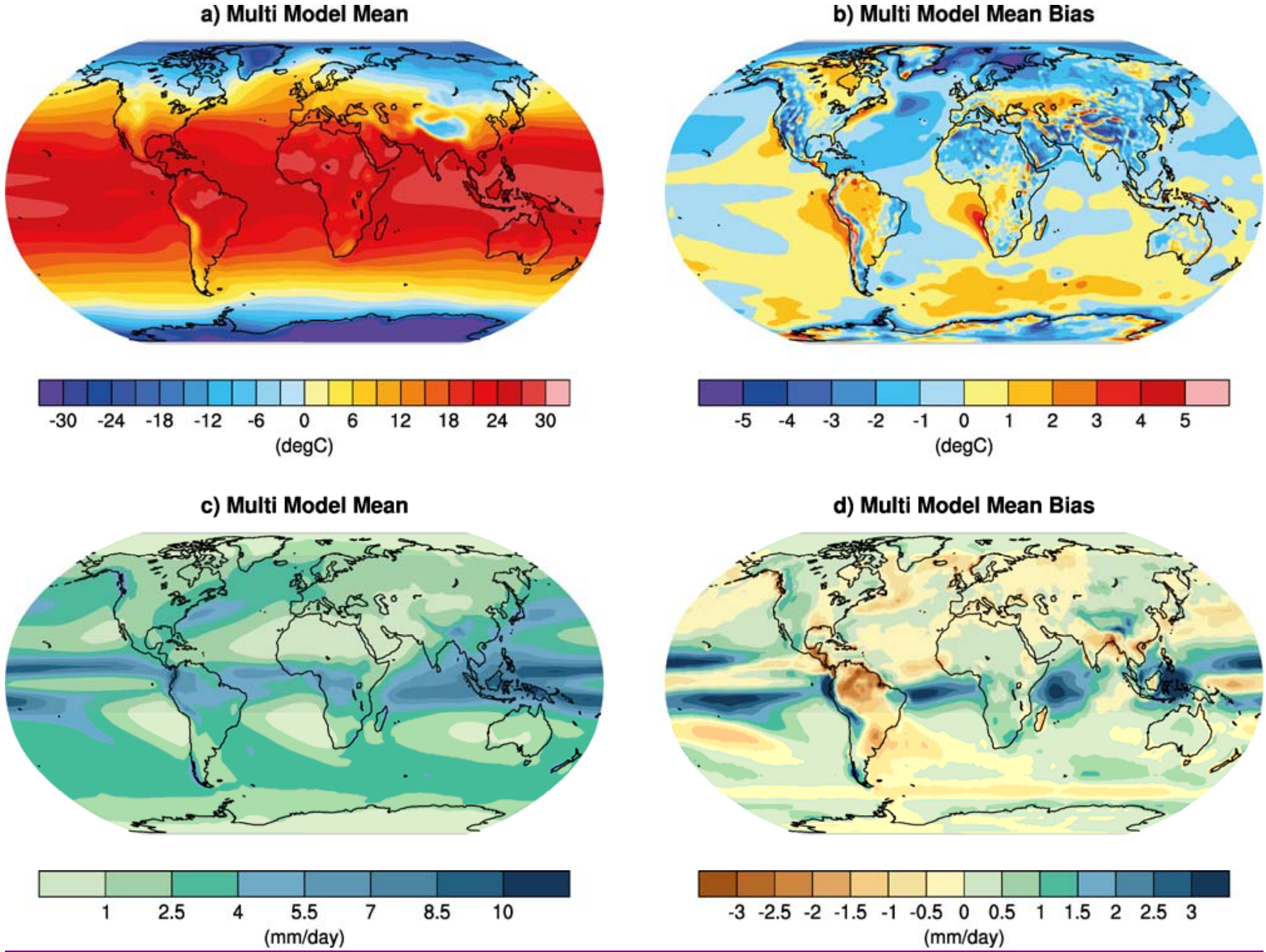


Figure 3. *Left.* Zonally averaged temperature profile difference between MPI-ESM-LR and the ERA-Interim reanalysis data with masked non-significant values. MPI-ESM-LR has generally small biases in the troposphere (< 1–~~2K~~2 K), but a cold bias in the tropopause region that is particularly strong in the extratropical lower stratosphere. This is a systematic bias present in many of the CMIP3 and CCMVal models (IPCC, 2007; SPARC-CCMVal, 2010), related to an overestimation of the water vapour concentrations in that region. *Right:* Taylor diagram for temperature at 850 hPa ~~for~~from CMIP5 models compared ~~to~~with ERA-Interim (reference observation-based data set) and NCEP (alternate observation-based data set) showing a very high correlation or R>0.98 with the reanalyses demonstrating very good performance in this quantity. Both figures produced with *namelist_perfmetrics_CMIP5.xml*.


a) Multi Model Mean
-30 -24 -18 -12 -6 0 6 12 18 24 30
(degC)
b) Multi Model Mean Bias
-5 -4 -3 -2 -1 0 1 2 3 4 5
(degC)
c) Multi Model Mean
1 2.5 4 5.5 7 8.5 10
(mm/day)
d) Multi Model Mean Bias
-3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 2.5 3
(mm/day)

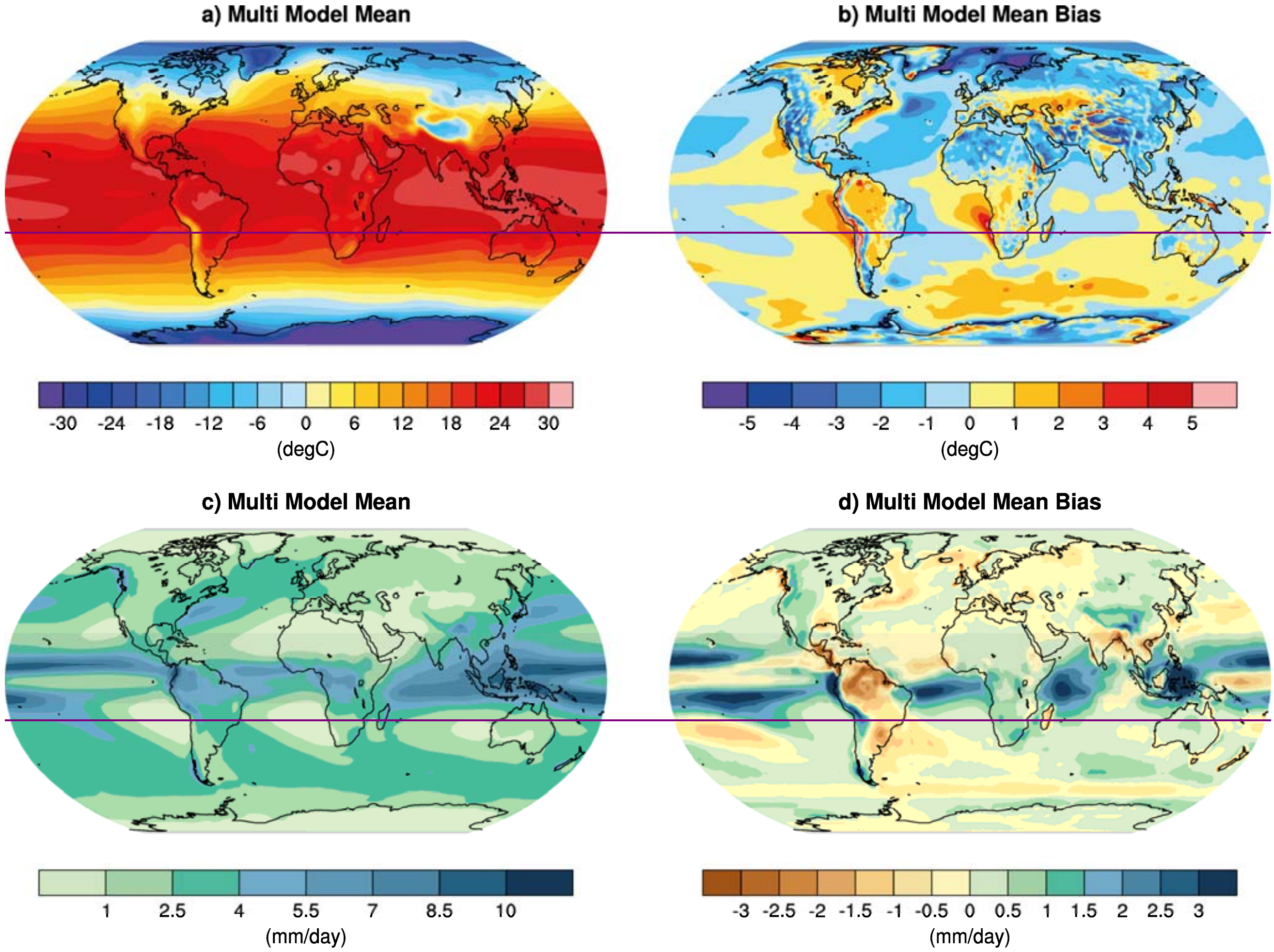
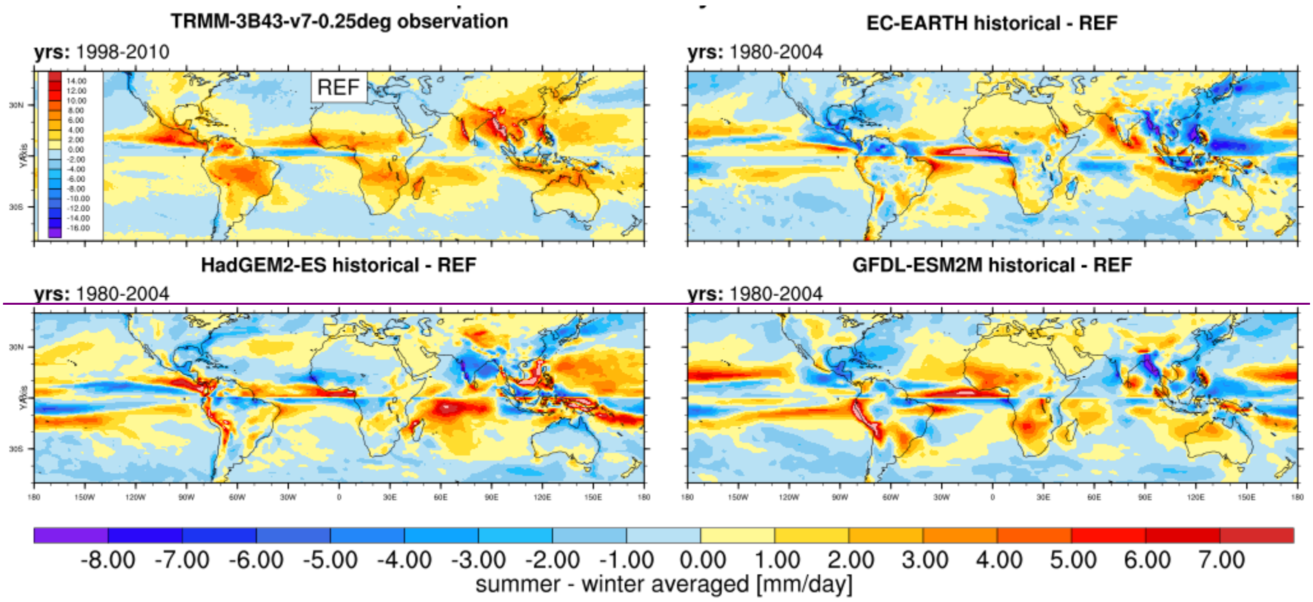
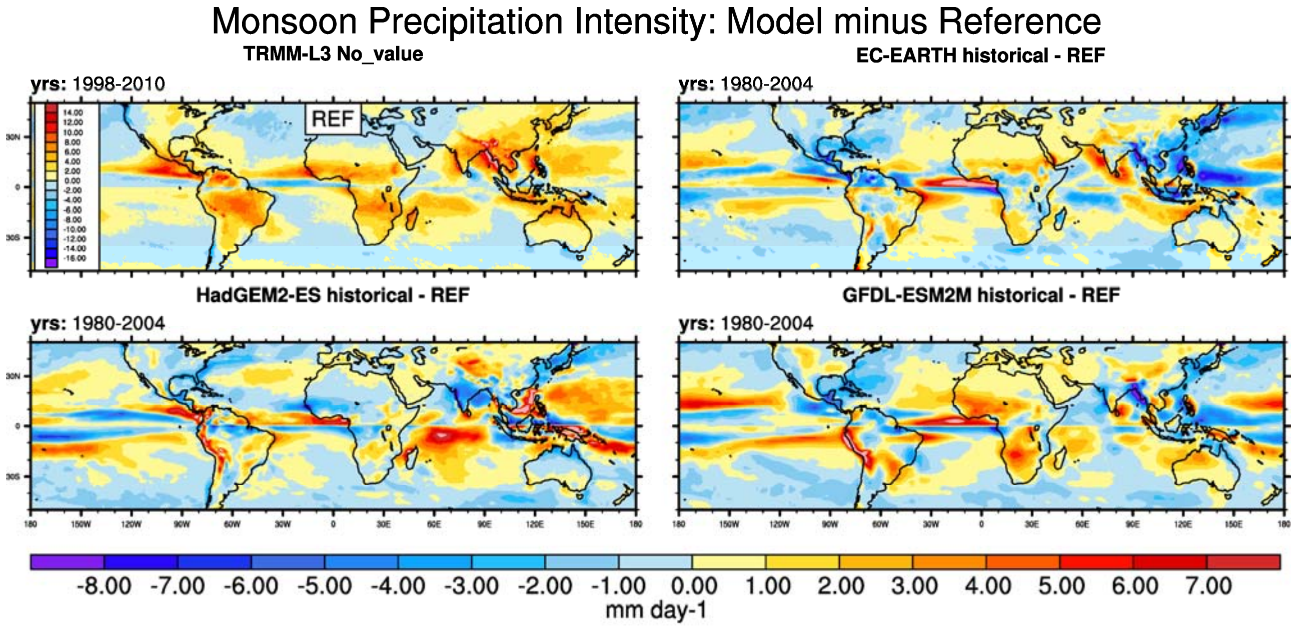
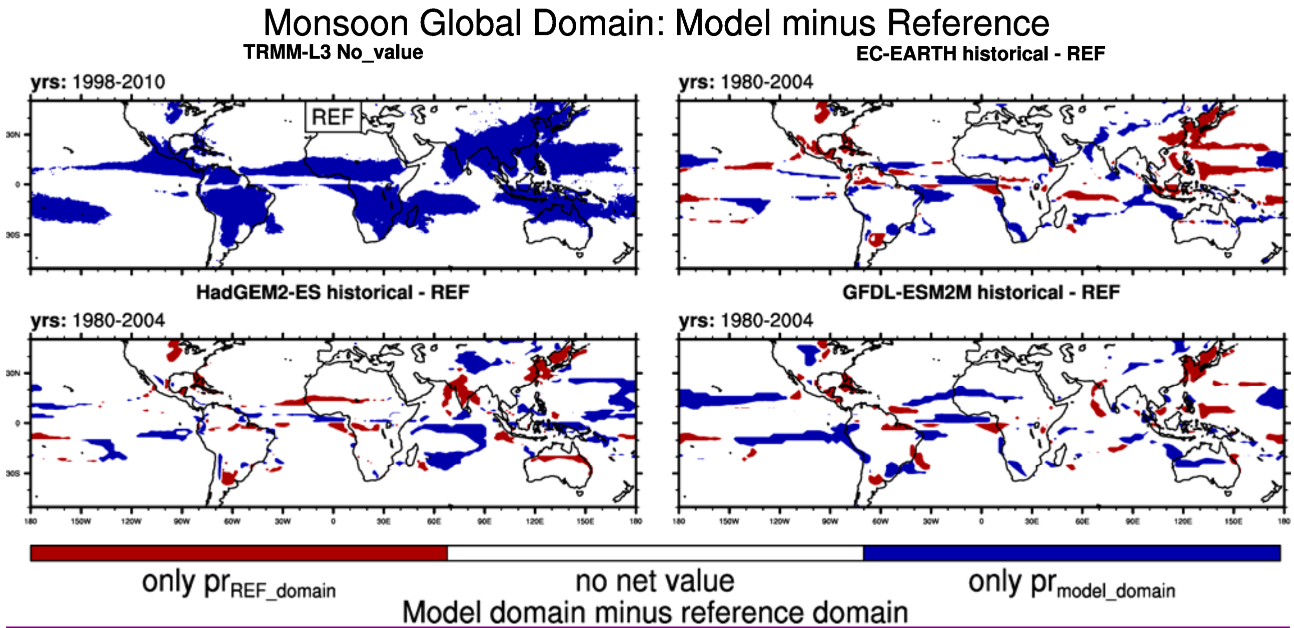


Figure 4. Annual-mean surface air temperature (upper row) and precipitation rate (mm day^{-1}) for the period 1980–2005. The left panels show the multi-model mean and the right panels the bias as the difference between the CMIP5 multi-model mean and the climatology from ERA-Interim (Dee et al., 2011) and the Global Precipitation Climatology Project (Adler et al., 2003) for surface air temperature and precipitation rate, respectively. The multi-model mean near-surface temperature agrees with ERA-Interim mostly within ±2°C. Larger biases can be seen in regions with sharp gradients in temperature, for example in areas with high topography such as the Himalaya, the sea ice edge in the North Atlantic, and over the coastal upwelling regions in the subtropical oceans. Biases in the simulated multi-model mean precipitation include too low precipitation along the equator in the western Pacific and too high precipitation amounts in the tropics south of the equator. ~~Similar to Figures 9.2 and 9.4 of Flato et al. (2013) and~~Similar to Figures 9.2 and 9.4 of Flato et al. (2013) and produced with *namelist_flato13ipcc.xml*.

1 
TRMM-3B43-v7-0.25deg observation
yrs: 1998-2010
REF
EC-EARTH historical - REF
yrs: 1980-2004
HadGEM2-ES historical - REF
yrs: 1980-2004
GFDL-ESM2M historical - REF
yrs: 1980-2004
-8.00 -7.00 -6.00 -5.00 -4.00 -3.00 -2.00 -1.00 0.00 1.00 2.00 3.00 4.00 5.00 6.00 7.00
summer - winter averaged [mm/day]


Monsoon Precipitation Intensity: Model minus Reference
TRMM-L3 No_value
EC-EARTH historical - REF
yrs: 1998-2010
REF
yrs: 1980-2004
HadGEM2-ES historical - REF
GFDL-ESM2M historical - REF
yrs: 1980-2004
yrs: 1980-2004
-8.00 -7.00 -6.00 -5.00 -4.00 -3.00 -2.00 -1.00 0.00 1.00 2.00 3.00 4.00 5.00 6.00 7.00
mm day-1


Monsoon Global Domain: Model minus Reference
TRMM-L3 No_value
EC-EARTH historical - REF
yrs: 1998-2010
REF
yrs: 1980-2004
HadGEM2-ES historical - REF
GFDL-ESM2M historical - REF
yrs: 1980-2004
yrs: 1980-2004
only pr_REF_domain
no net value
only pr_model_domain
Model domain minus reference domain

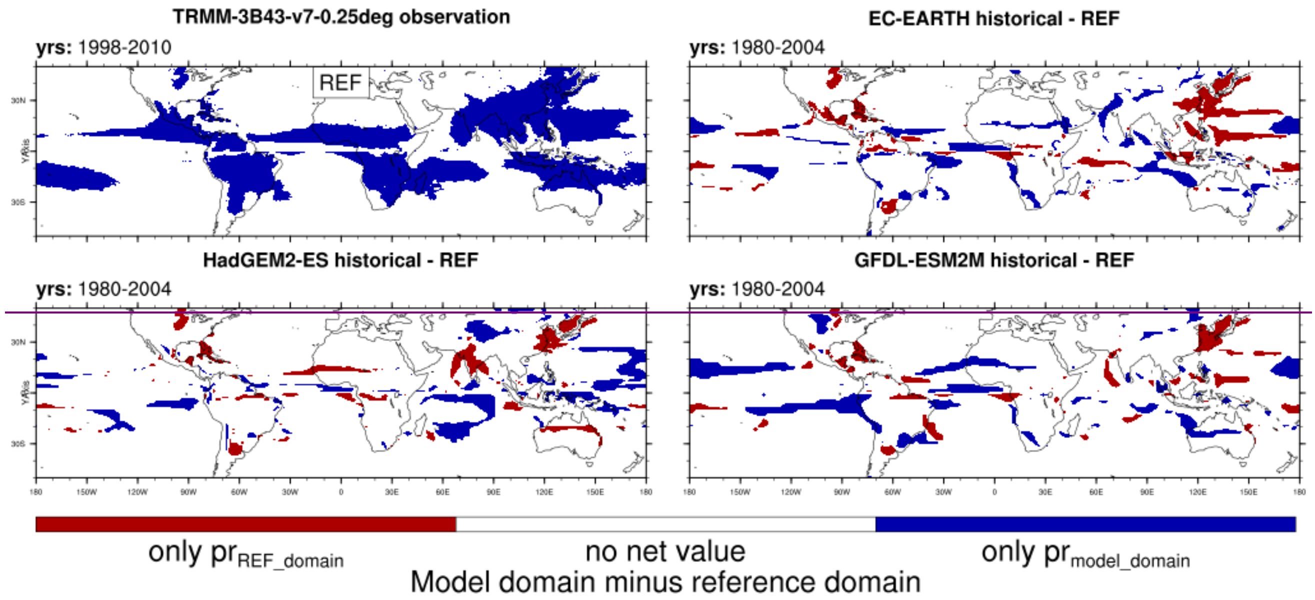
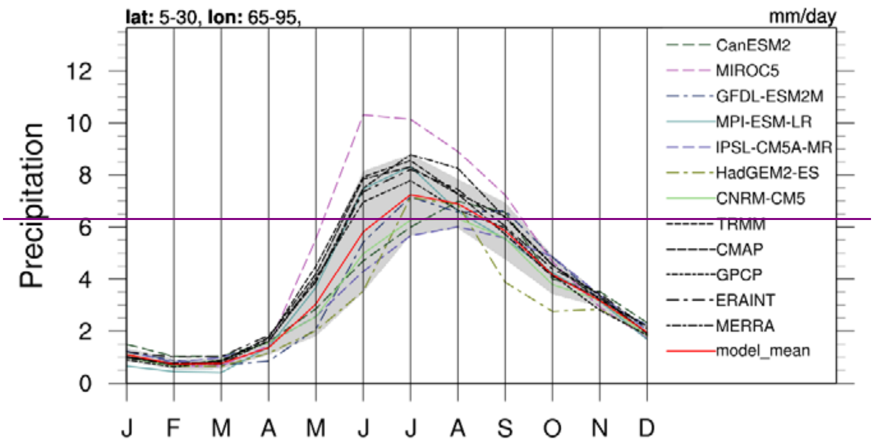
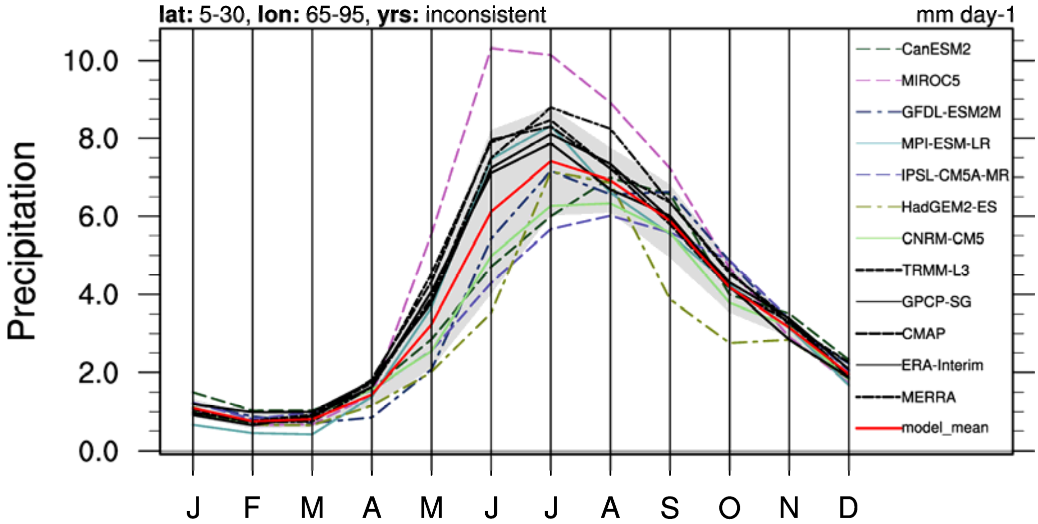


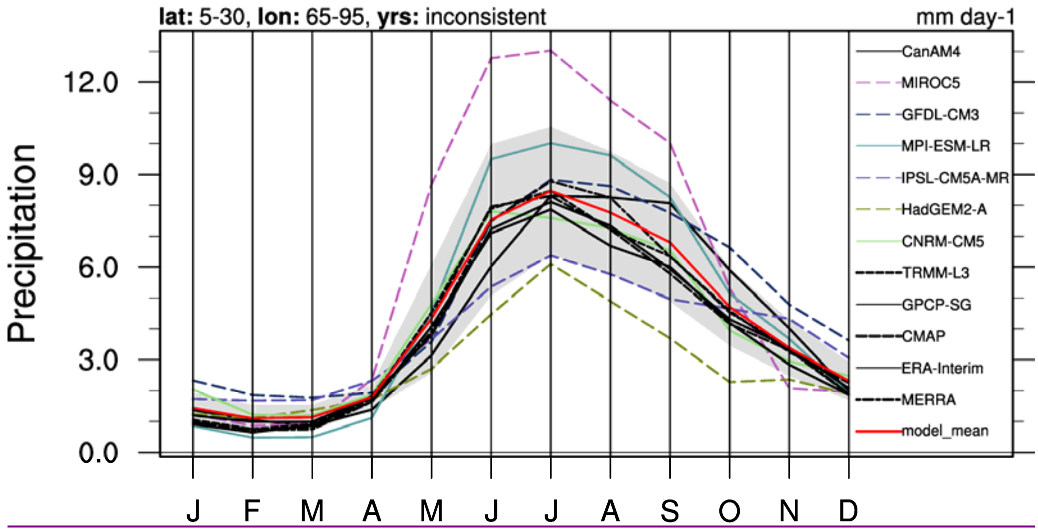
Figure 5. Monsoon precipitation intensity (upper panels) and monsoon precipitation domain (lower panels) for TRMM and an example of deviations from observations from three CMIP5 models (EC-Earth, HadGEM2-ES, and GFDL-ESM2M). ~~Models~~The models have difficulties representing the eastward extent of the monsoon domain over the South China Sea and western Pacific, and several models (e.g., HadGEM2-ES) underestimate the latitudinal extent of most of the monsoon regions. The monsoon precipitation intensity tends to be underestimated in the South Asian, East Asian and Australian monsoon regions while in the African and American monsoon regions the sign of the intensity bias varies between models. ~~Similar to Figure 9.32 of Flato et al. (2013)~~Similar to Figure 9.32 of Flato et al. (2013) and produced with *namelist_SAMonsoon.xml*.



Seasonal cycle



Seasonal cycle



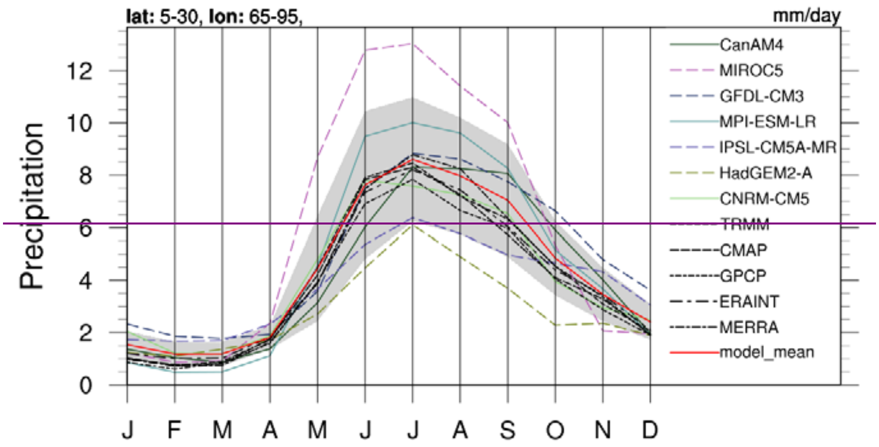


Figure 6. Seasonal cycle of monthly rainfall averaged over the Indian region (5-~~30N~~30°N, 65-~~95E~~95°E) for a range of CMIP5 coupled models (upper panel) and their AMIP counterparts (lower panel), averaged over available years (models: 1980-2004, observations: 1998-~~2009~~2010). The grey area in each panel indicates standard deviation from the model mean, to indicate the spread between models (observations/reanalyses are not included in this spread). These illustrate the range of rainfall simulated particularly in AMIP experiments where there is no feedback between precipitation and SST biases that might moderate the rainfall biases (Bollasina and Ming, 2013; Levine et al., 2013). Some of the CMIP5 coupled models (e.g., HadGEM2-ES, IPSL-CM5A-MR) show a delayed monsoon onset that is not apparent in their AMIP configurations. This is related to cold SST biases in the Arabian Sea which develop during boreal winter and spring (Levine et al., 2013). Produced with *namelist_SAMonsoon.xml*.

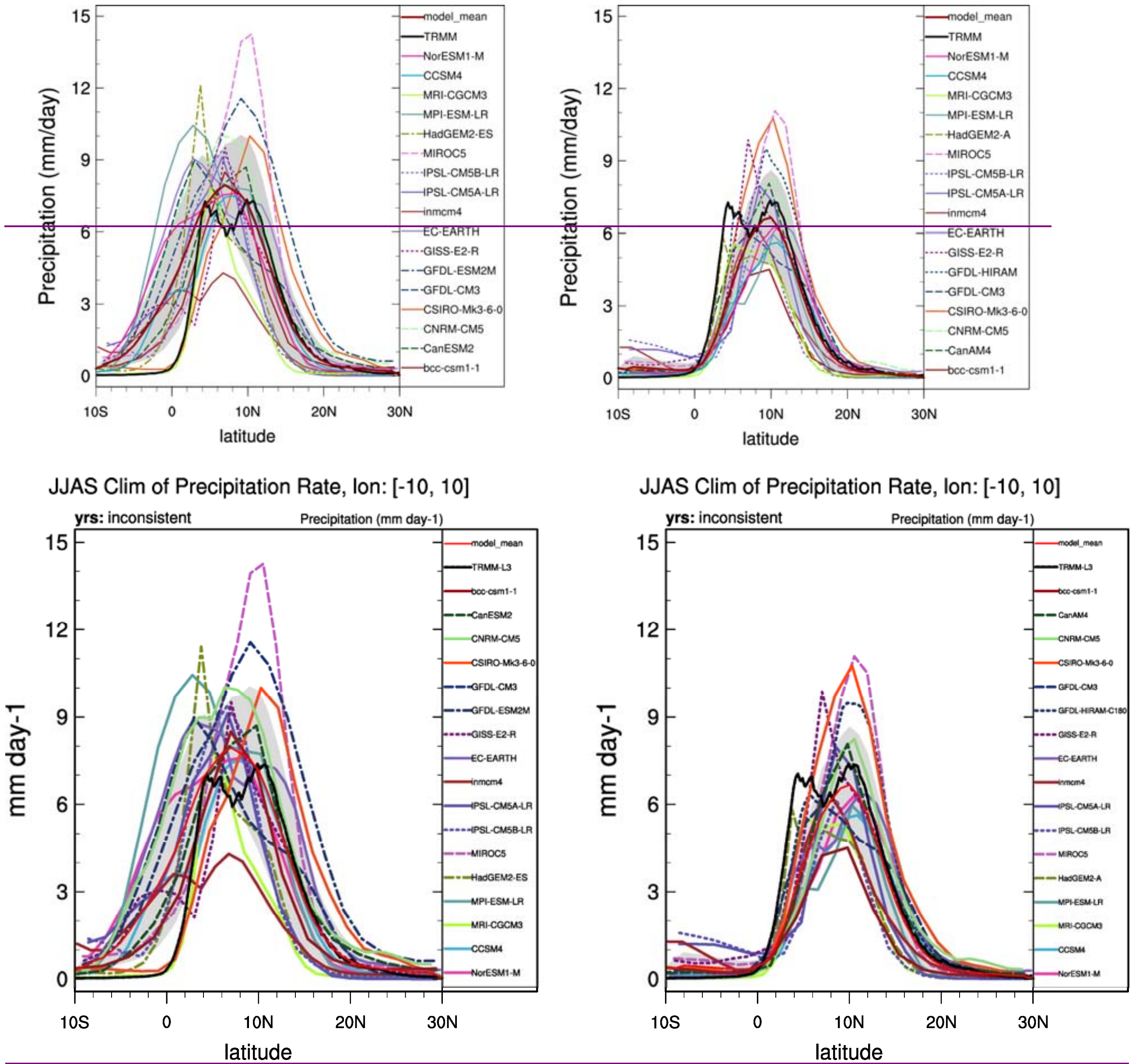
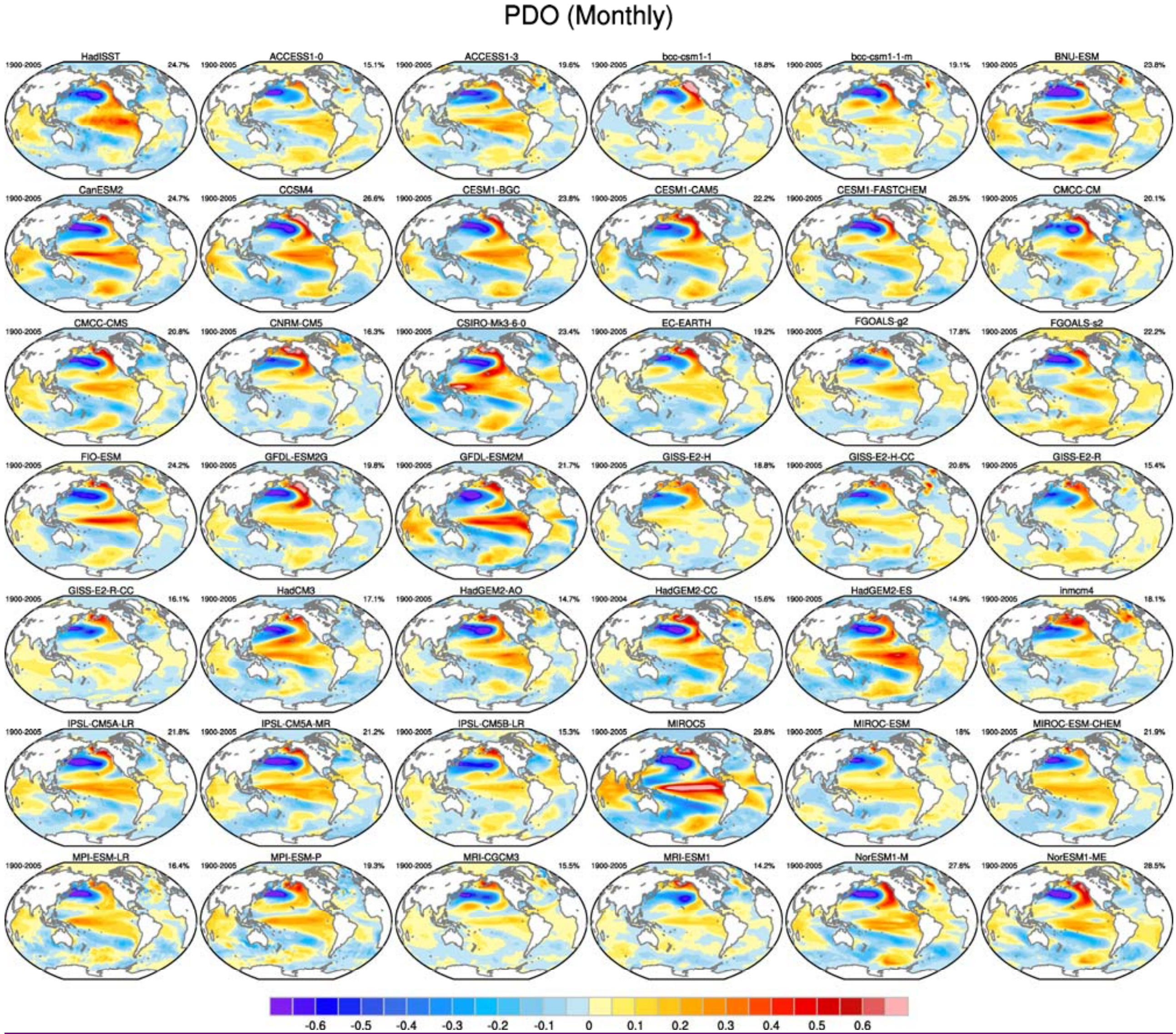


Figure 7. Precipitation (mm day^{-1}) averaged over 10°W-10°E for the JJAS season for the years 1979-2005 for CMIP5 historical simulations (left) and 1979-2008 for CMIP5 AMIP simulations (right) compared to 1998-2008 for TRMM 3B43 Version 7 data set. The results illustrate the inter-model spread in the mean position and intensity of the WAM among the CMIP5 models. The spread is slightly reduced in AMIP simulations, as the warm SST bias in the equatorial Atlantic is removed. The WAM mean structure, however, is not captured by many models. Produced with *namelist_WAMonsoon.xml*.


PDO (Monthly)
1900-2005 HadISST 24.7%
1900-2005 ACCESS1-0 15.1%
1900-2005 ACCESS1-3 19.6%
1900-2005 bcc-csm1-1 18.8%
1900-2005 bcc-csm1-1-m 19.1%
1900-2005 BNU-ESM 23.8%
1900-2005 CanESM2 24.7%
1900-2005 CCSM4 26.6%
1900-2005 CESM1-BGC 23.8%
1900-2005 CESM1-CAM5 22.2%
1900-2005 CESM1-FASTCHEM 26.5%
1900-2005 CMCC-CM 20.1%
1900-2005 CMCC-CMS 20.8%
1900-2005 CNRM-CM5 16.3%
1900-2005 CSIRO-Mk3-6-0 23.4%
1900-2005 EC-EARTH 19.2%
1900-2005 FGOALS-g2 17.8%
1900-2005 FGOALS-s2 22.2%
1900-2005 FIO-ESM 24.2%
1900-2005 GFDL-ESM2G 19.8%
1900-2005 GFDL-ESM2M 21.7%
1900-2005 GISS-E2-H 18.8%
1900-2005 GISS-E2-H-CC 20.6%
1900-2005 GISS-E2-R 15.4%
1900-2005 GISS-E2-R-CC 16.1%
1900-2005 HadCM3 17.1%
1900-2005 HadGEM2-AO 14.7%
1900-2004 HadGEM2-CC 15.6%
1900-2005 HadGEM2-ES 14.9%
1900-2005 inmcm4 18.1%
1900-2005 IPSL-CM5A-LR 21.8%
1900-2005 IPSL-CM5A-MR 21.2%
1900-2005 IPSL-CM5B-LR 15.3%
1900-2005 MIROC5 29.8%
1900-2005 MIROC-ESM 18%
1900-2005 MIROC-ESM-CHEM 21.9%
1900-2005 MPI-ESM-LR 16.4%
1900-2005 MPI-ESM-P 19.3%
1900-2005 MRI-CGCM3 15.5%
1900-2005 MRI-ESM1 14.2%
1900-2005 NorESM1-M 27.8%
1900-2005 NorESM1-ME 28.5%
-0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0 0.1 0.2 0.3 0.4 0.5 0.6

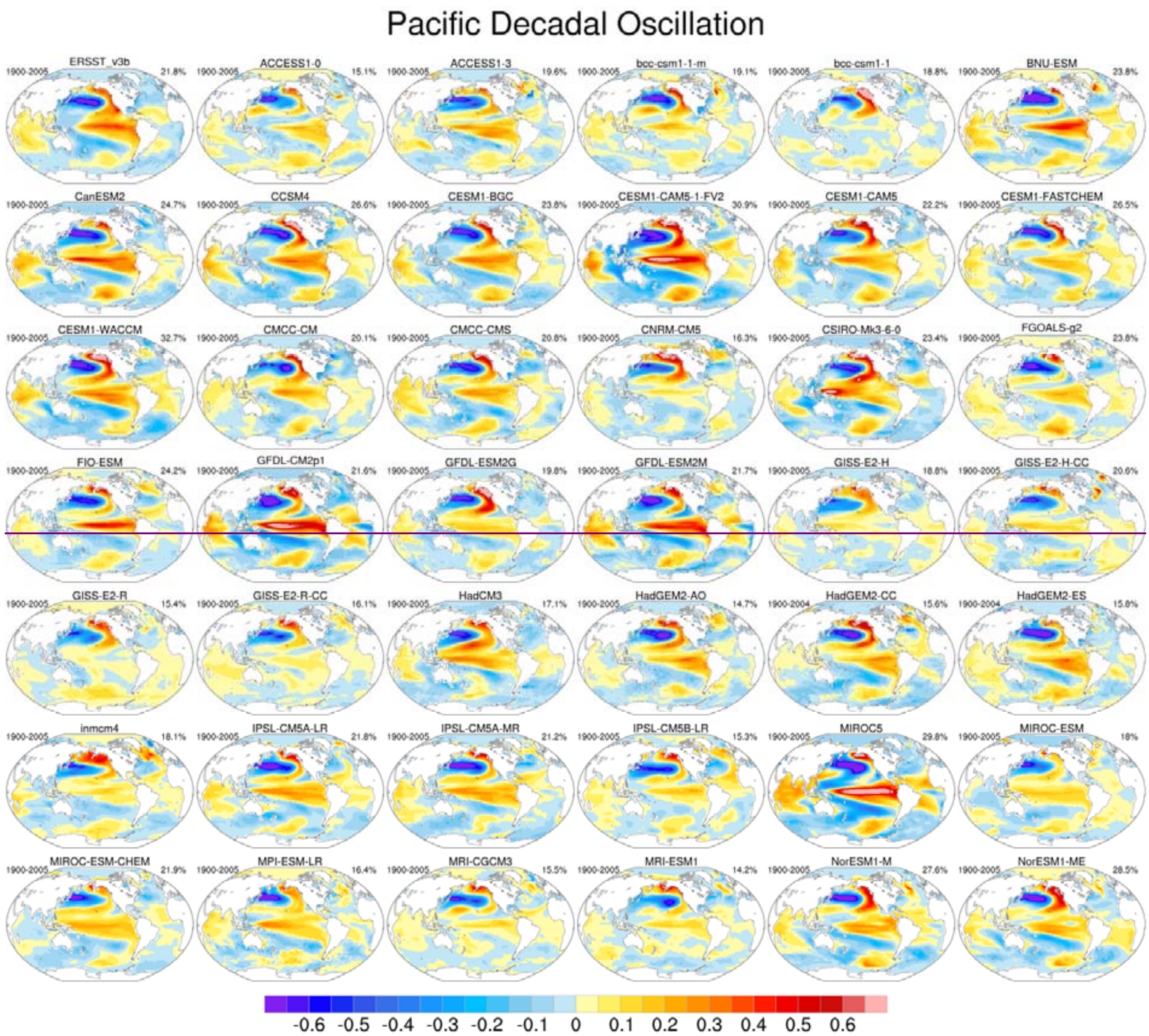
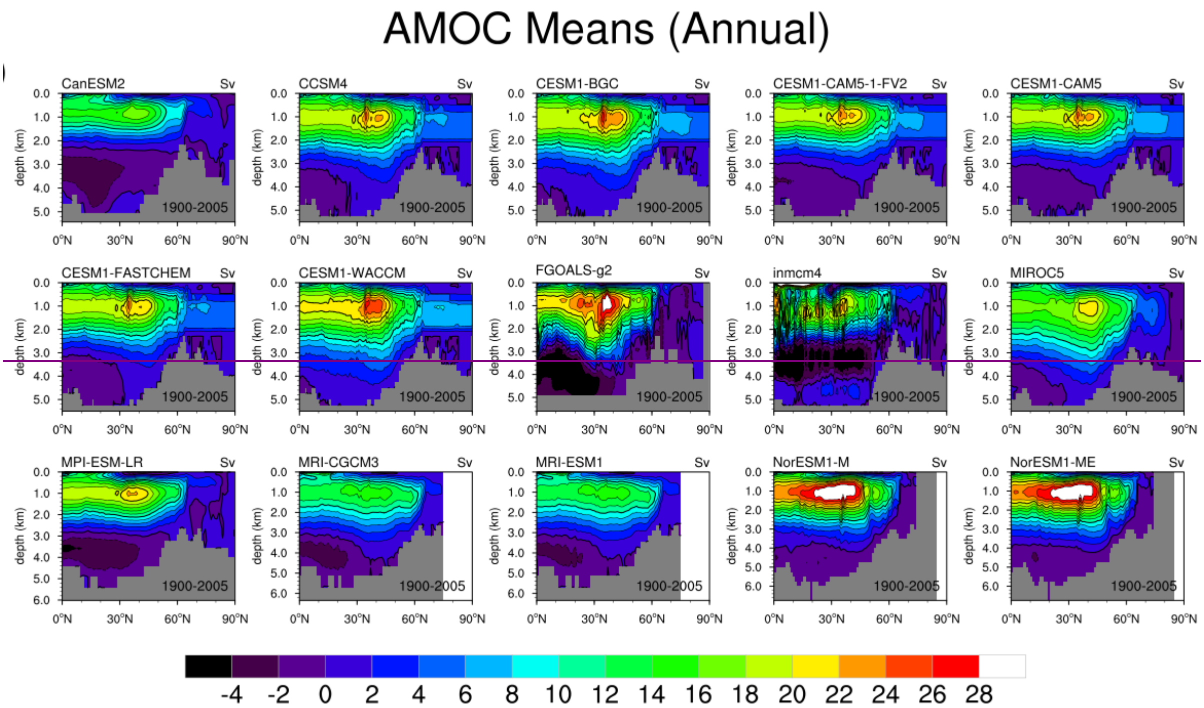


Figure 8. The PDO as simulated by 41 CMIP5 models (individual panels labelled by model name) and observations (upper left panel) for the historical period 1900-2005. These patterns show the global SST anomalies (°C) associated with a one standard deviation change in the normalized principal component (PC) time series. The percent variance accounted by the PDO is given in the upper right of each panel. The PDO is defined as the leading empirical orthogonal function of monthly SST anomalies (minus the global mean SST) over the North Pacific (20-70°N, 110°E-100°W). The global patterns (°C) are formed by regressing monthly SST anomalies at each grid point onto the PC time series. Most CMIP5 models show realistic patterns in the North Pacific. However, linkages with the tropics and the tropical Pacific in particular, vary across models. The lack of a strong tropical expression of the PDO is a major shortcoming in many CMIP5 models (Flato et al., 2013). Figure produced with *namelist_CVDP.xml*.


AMOC Means (Annual)
CanESM2
CCSM4
CESM1-BGC
CESM1-CAM5-1-FV2
CESM1-CAM5
CESM1-FASTCHEM
CESM1-WACCM
FGOALS-g2
inmcm4
MIROC5
MPI-ESM-LR
MRI-CGCM3
MRI-ESM1
NorESM1-M
NorESM1-ME
Sv
depth (km)
1900-2005
0°N
30°N
60°N
90°N
-4 -2 0 2 4 6 8 10 12 14 16 18 20 22 24 26 28

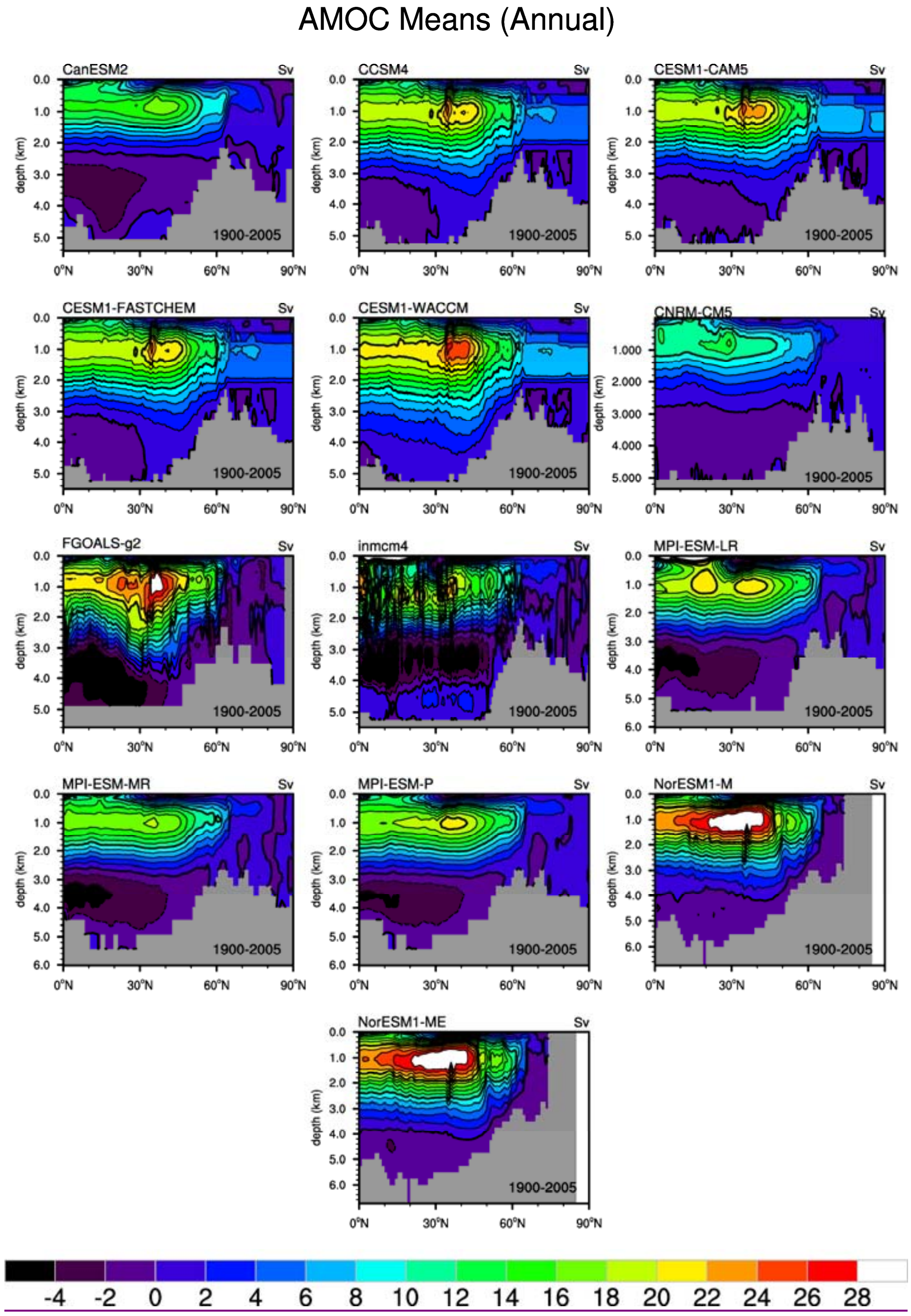
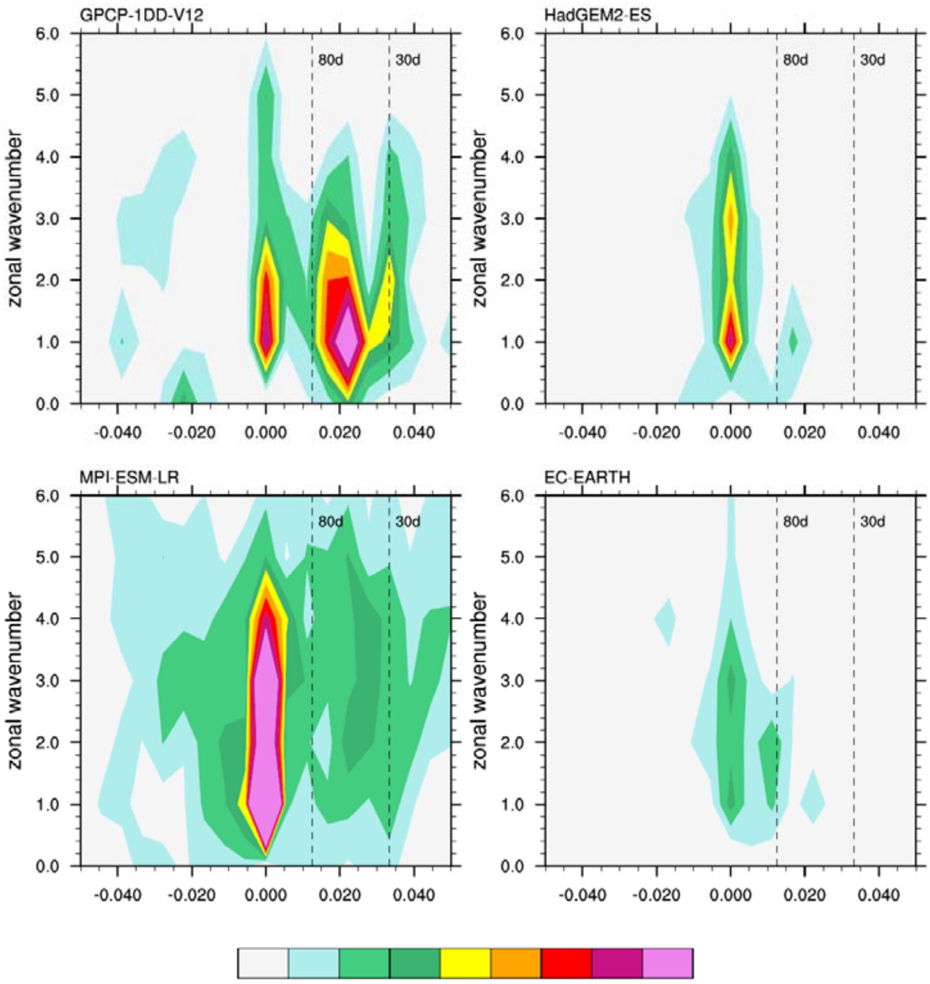


Figure 9. Long-term annual mean Atlantic Meridional Overturning Streamfunction (AMOC; Sv) as simulated by ~~15~~13 CMIP5 models (individual panels labelled by model name) for the historical period 1900-2005. ~~MOC~~AMOC annual averages are formed, weighted by the cosine of the latitude and by the depth of the vertical layer, and then the data is masked by setting all those areas to missing where the variance is less than 1.e^{-6}. The figure shows that there is a wide spread among the CMIP5 models, with maximal AMOC strength ranging from ~13 Sv (CanESM2) to over ~28 Sv (NorESM1), while the models agree generally well on the position of maximal AMOC strength. Figure produced with *namelist_CVDP.xml*.

1


GPCP-1DD-V12
HadGEM2-ES
MPI-ESM-LR
EC-EARTH
zonal wavenumber
80d
30d
0.01 0.015 0.02 0.025 0.03 0.035 0.04 0.05

summer

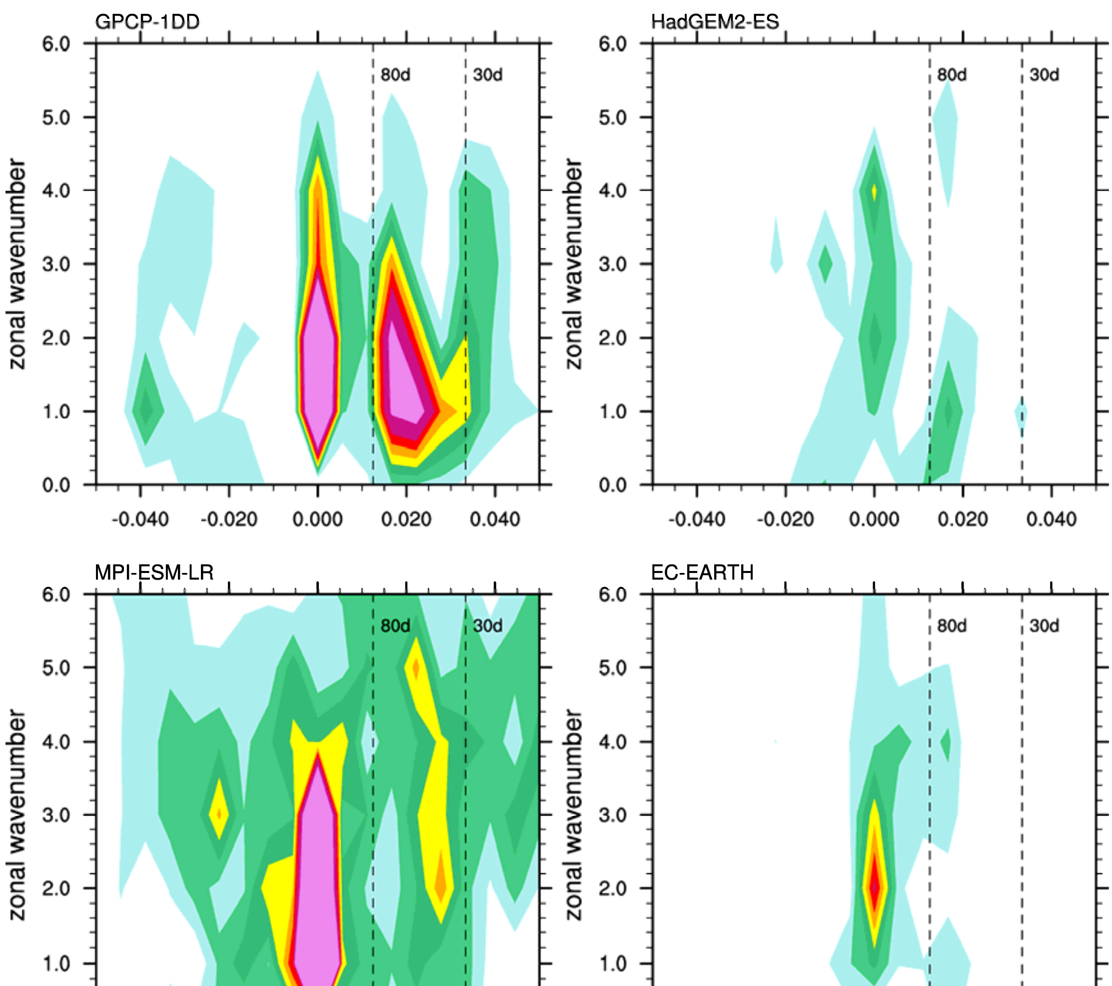
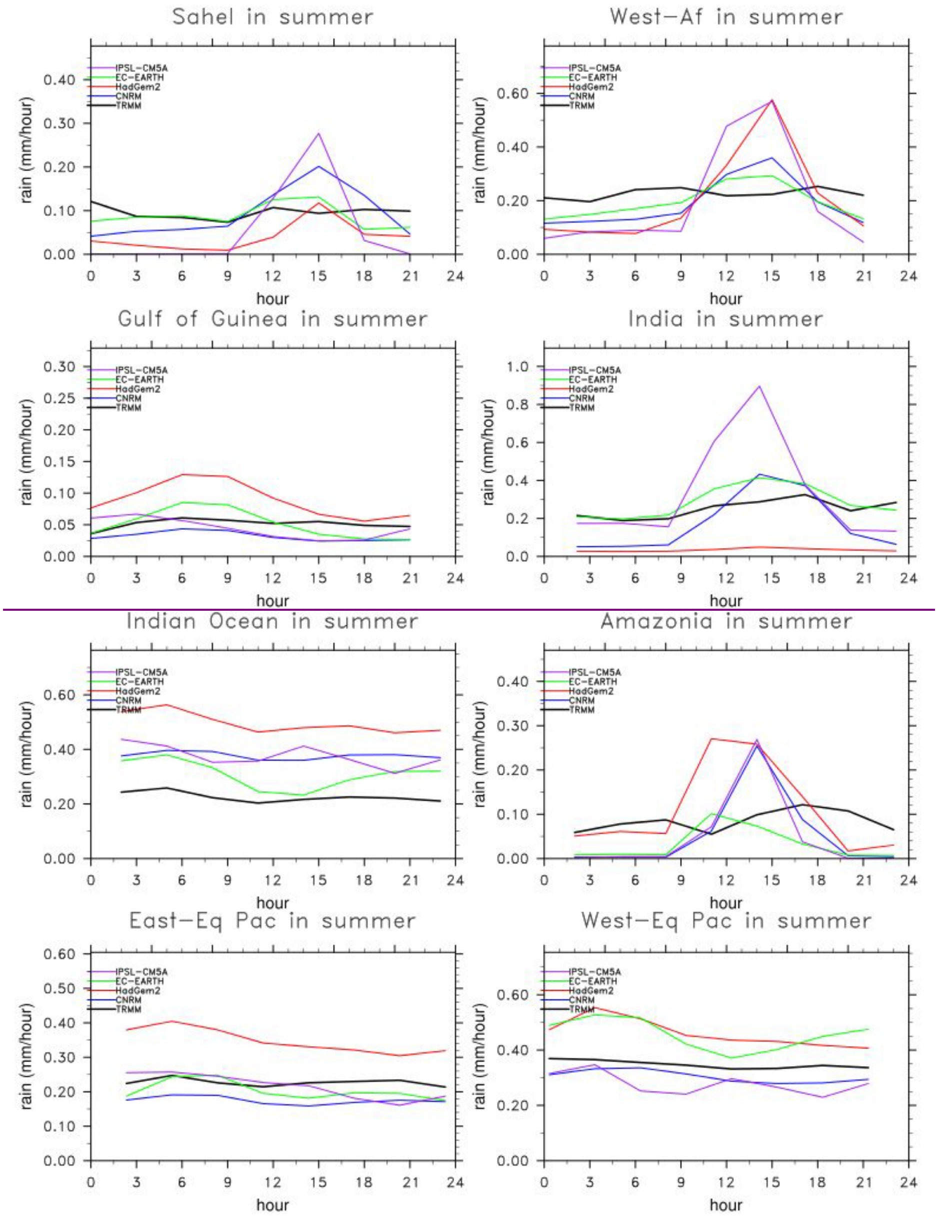

GPCP-1DD
HadGEM2-ES
MPI-ESM-LR
EC-EARTH
zonal wavenumber
80d
30d

Figure 10. May-October wavenumber-frequency spectra of 10ºS-10ºN averaged precipitation (mm^2 day^{-2}) for GPCP-1DD, ~~HadGEM~~HadGEM2-ES, MPI-ESM-LR and EC-~~EARTH~~Earth. Individual May-October spectra were calculated for each year and then averaged over all years of data. Only the climatological seasonal cycle and time mean for each May-October segment were removed before calculation of the spectra. The bandwidth is (180 days)$^{-1}$. The observed precipitation shows the dominant MJO spatial scale is zonal wavenumber 1-3 at the 30-~~80day~~80-day frequency. According to the definition, the positive frequency ~~represent~~represents eastward propagation of the MJO. Compared with observations, both ~~HadGEM~~HadGEM2-ES and EC-~~EARTH~~Earth models have difficulties simulating precipitation variability on MJO timescsales. Produced with *namelist_mjo_daily.xml.*


Sahel in summer
West-Af in summer
Gulf of Guinea in summer
India in summer
Indian Ocean in summer
Amazonia in summer
East-Eq Pac in summer
West-Eq Pac in summer
IPSL-CM5A
EC-EARTH
HadGem2
CNRM
TRMM
rain (mm/hour)
hour

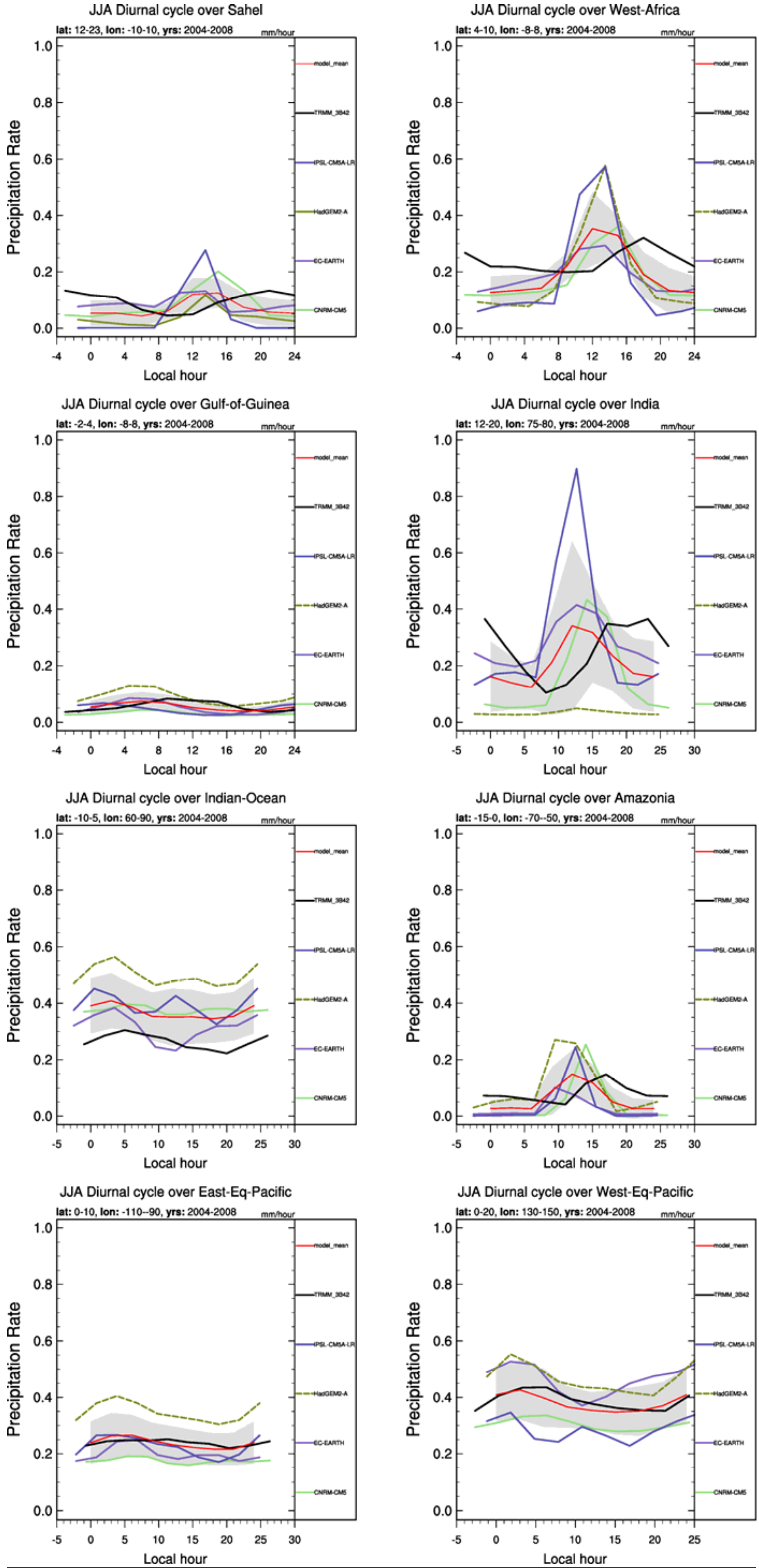
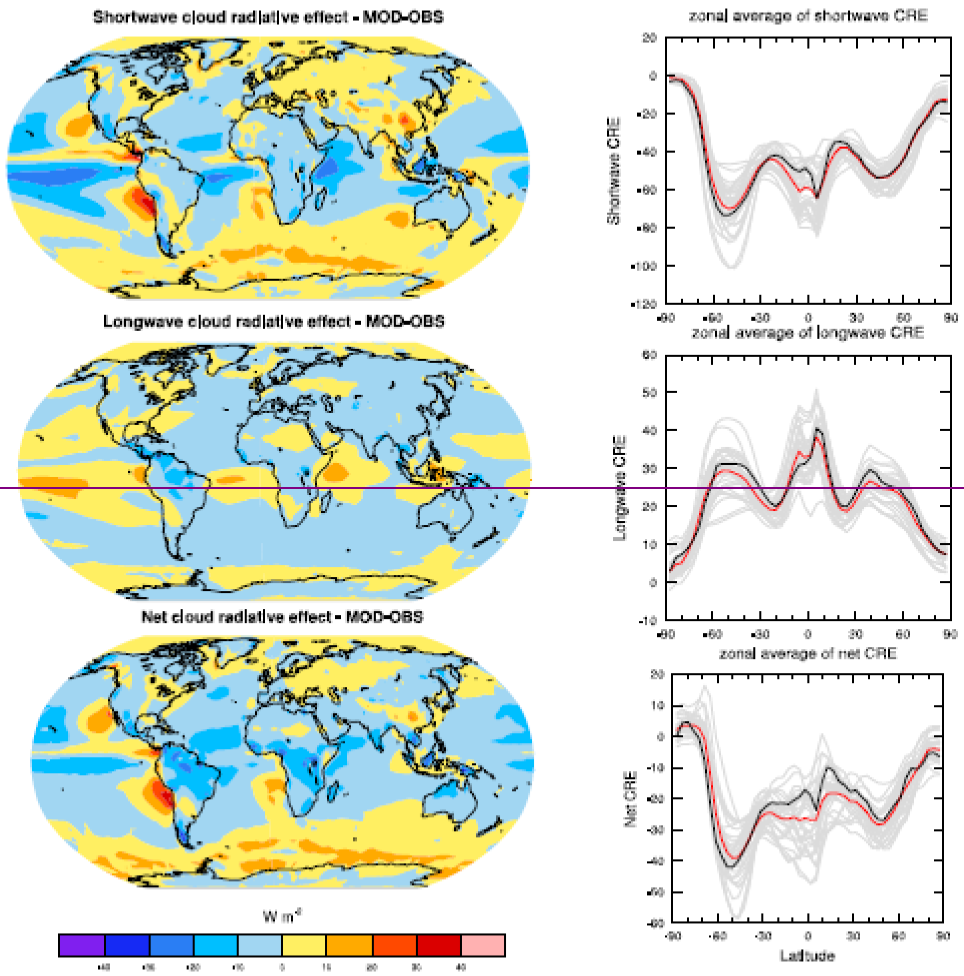

JJA Diurnal cycle over Sahel
lat: 12-23, lon: -10-10, yrs: 2004-2008
JJA Diurnal cycle over West-Africa
lat: 4-10, lon: -8-8, yrs: 2004-2008
JJA Diurnal cycle over Gulf-of-Guinea
lat: -2-4, lon: -8-8, yrs: 2004-2008
JJA Diurnal cycle over India
lat: 12-20, lon: 75-80, yrs: 2004-2008
JJA Diurnal cycle over Indian-Ocean
lat: -10-5, lon: 60-90, yrs: 2004-2008
JJA Diurnal cycle over Amazonia
lat: -15-0, lon: -70--50, yrs: 2004-2008
JJA Diurnal cycle over East-Eq-Pacific
lat: 0-10, lon: -110--90, yrs: 2004-2008
JJA Diurnal cycle over West-Eq-Pacific
lat: 0-20, lon: 130-150, yrs: 2004-2008
mm/hour
Precipitation Rate
Local hour
model_mean
TRMM_3B42
IPSL-CM5A-LR
HadGEM2-A
EC-EARTH
CNRM-CM5

Figure 11. Mean diurnal cycle of precipitation (mm/hour) averaged over five summers (2004-2008) over specific regions in the tropics (Sahel, West-Africa, Gulf of Guinea, India, Indian Ocean, Amazonia, East-Equatorial Pacific and West-Equatorial Pacific) as observed by TRMM 3B42 ~~V6~~V7 and as simulated by four CMIP5 models: CNRM-CM5, EC-Earth, ~~HadGem2~~HadGEM2-A and IPSL-CM5A-LR. ESMs produce a too strong peak of rainfall around noon over land while the observed precipitation maximum is weaker and delayed to 6 pm. At the same time, most models underestimate nocturnal precipitation. Over the ocean, the diurnal cycle of precipitation is more flat but rainfall maximum usually occurs a few hours earlier than in observations during the night, and the amplitude of oceanic precipitation shows large variations among models. Produced with *namelist_~~diurnal~~DiurnalCycle_box_pr.xml*.


Shortwave cloud radiative effect - MOD-OBS
Longwave cloud radiative effect - MOD-OBS
Net cloud radiative effect - MOD-OBS
W m⁻²
zonal average of shortwave CRE
Shortwave CRE
zonal average of longwave CRE
Longwave CRE
zonal average of net CRE
Net CRE
Latitude

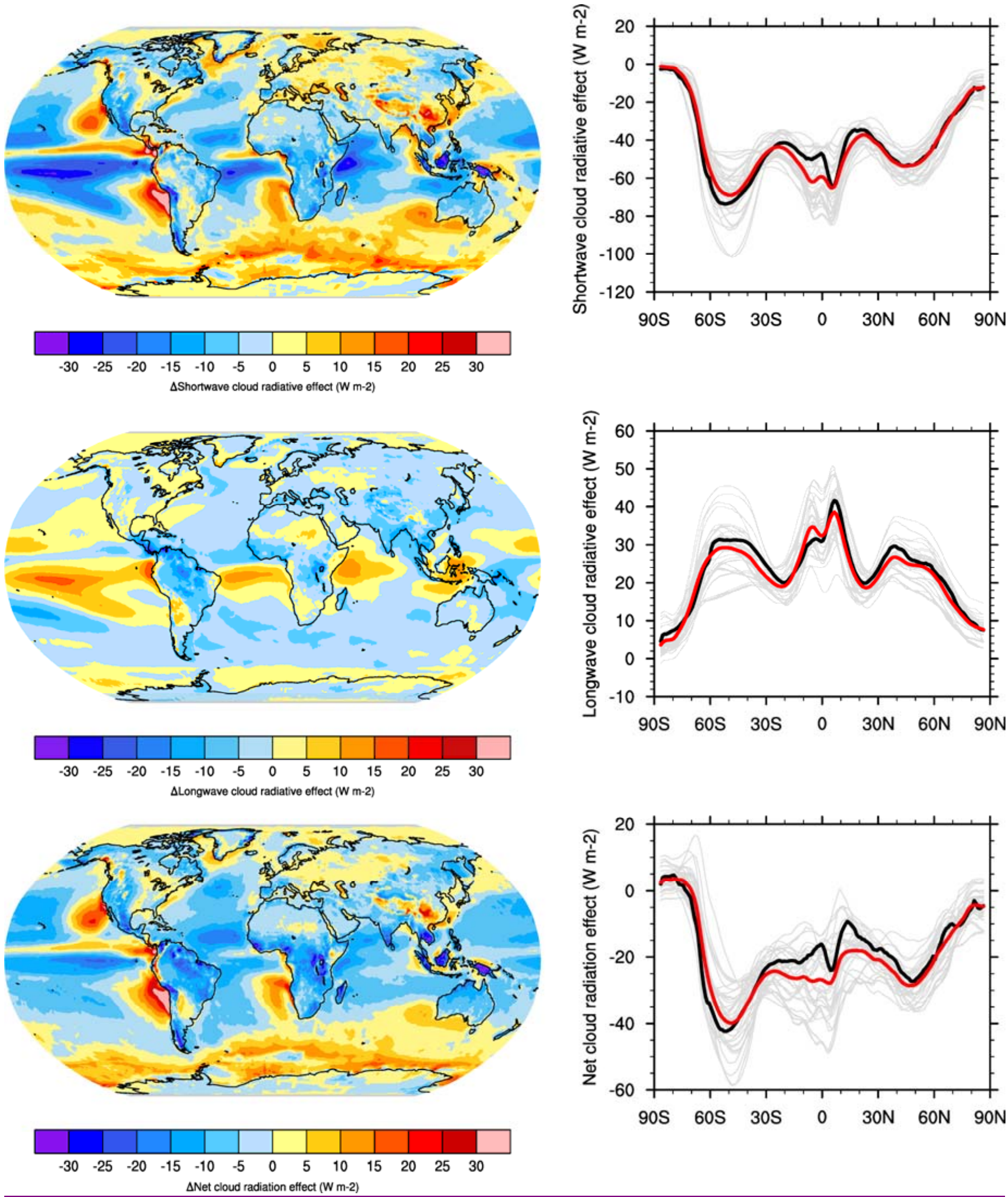
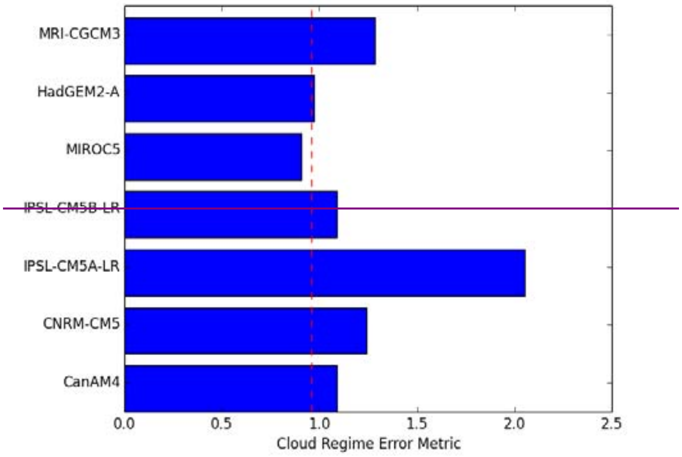
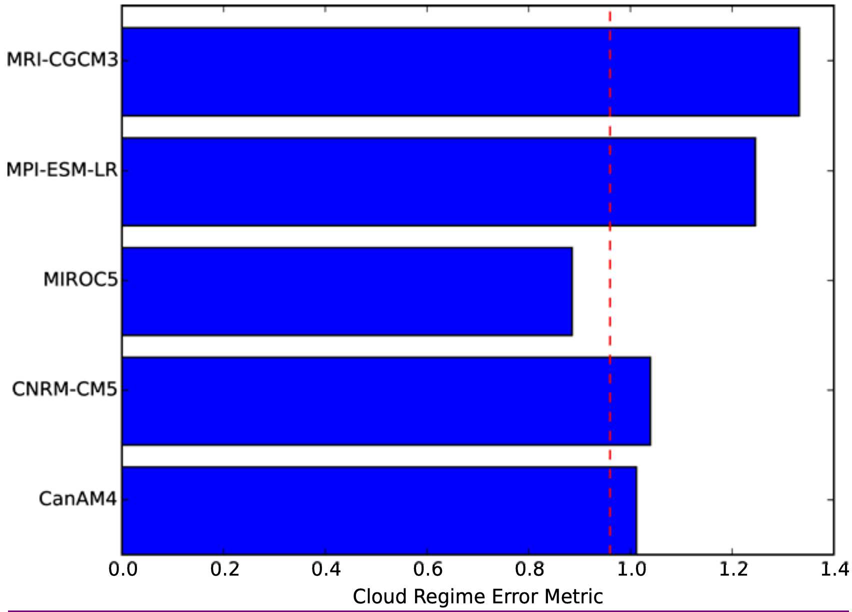


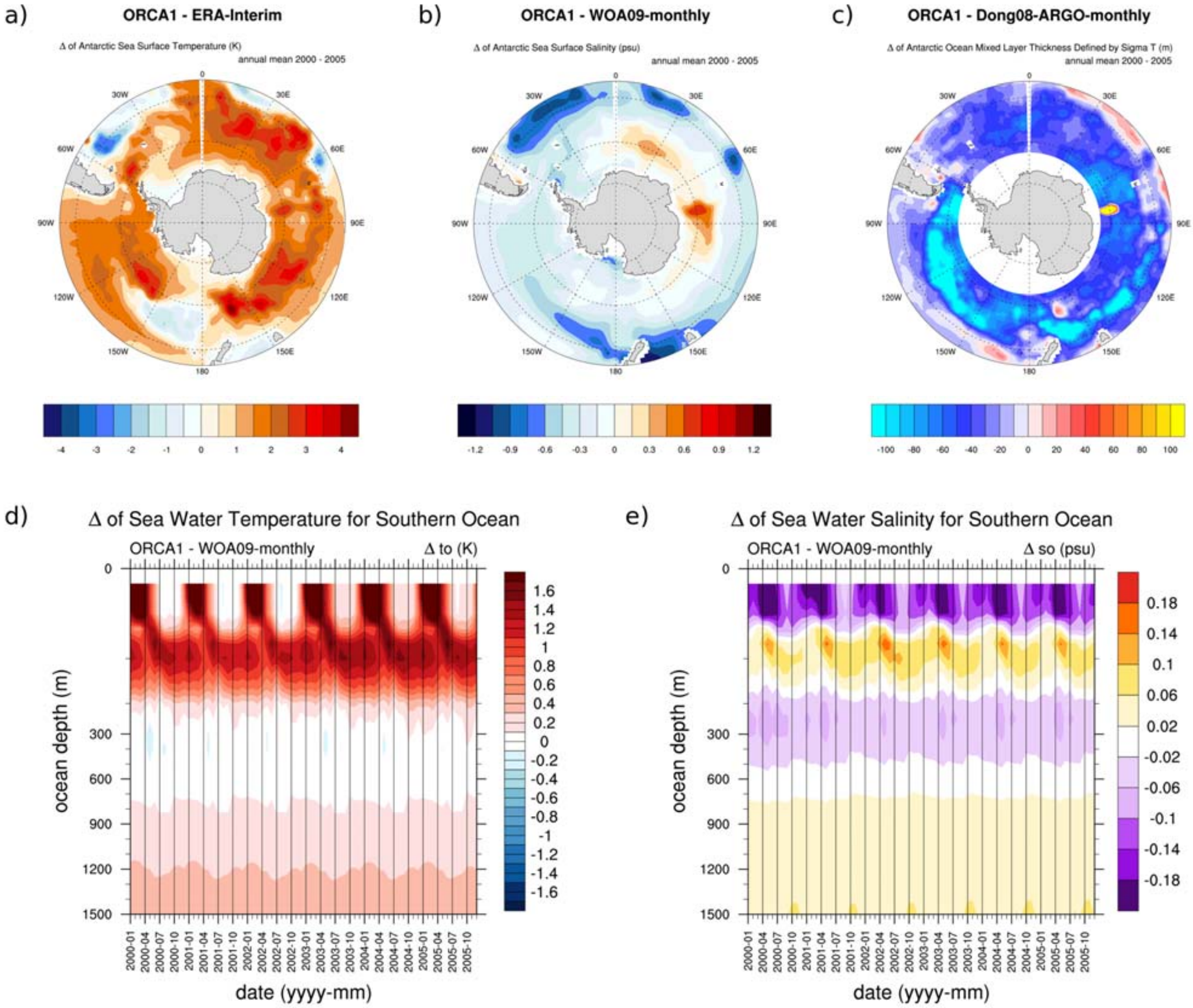
Figure 12. Climatological (1985-2005) annual-mean cloud radiative effects ~~in Wm^{-2} for~~from the CMIP5 models against CERES EBAF (2001–~~2011~~2012) in W m^{-2}. Top row shows the shortwave effect; middle row the longwave effect, and bottom row the net effect. Multi-model-mean biases against CERES EBAF 2.~~6~~7 are shown on the left, whereas the right panels show zonal averages from CERES EBAF 2.~~6 (dashed black), CERES ES-4~~7 (black), the individual CMIP5 models (thin grey lines), and the multi-model mean (thick red line). The multi-model mean longwave CRE is

overestimated in models, particularly in the Pacific and Atlantic south of the inter-tropical convergence zone (ITCZ) and in the South Pacific convergence zone (SPCZ). The longwave CRE is underestimated over Central and South America as well as parts of Central Africa and southern Asia. The most striking biases in the multi-model mean shortwave CRE are found in the stratocumulus regions off the west coasts of North and South America, southern Africa, and Australia. Despite biases in component cloud properties, simulated CRE is in quite good agreement with observations. ~~Reproducing Figure 9.5 of Flato et al. (2013)~~Reproducing Figure 9.5 of Flato et al. (2013) and produced with *namelist_flato13ipcc.nml*.





~~Figure 13. Cloud Regime Error Metric (CREM) from Williams and Webb (2009)~~Figure 13. Cloud Regime Error Metric (CREM) from Williams and Webb (2009) applied to ~~those~~some CMIP5 AMIP simulations with the required data in the archive. The results show that MIROC5 is the best performing model on this metric~~ with HadGEM2-A also having a score comparable to the observational uncertainty. Other~~, other models are slightly worse ~~and IPSL-CM5A-LR is notably deficient~~ on this metric. The red dashed line shows the observational uncertainty estimated from applying this metric to independent data from MODIS. An advantage of the metric is that its components can be decomposed to investigate the reasons for poor performance. This requires extra print statements compared to the default code~~,~~ but ~~when this is done~~might help to identify, for ~~IPSL-CM5A-LR it is found that a number of the~~instance, cloud regimes that are too reflective ~~(e.g. extra-tropical shallow cumulus and transition regimes) along with the stratocumulus regime being~~or simulated too frequently at the expense of some of the other regimes. Produced with *namelist_williams09climdyn_CREM.xml*.


a)
ORCA1 - ERA-Interim
Δ of Antarctic Sea Surface Temperature (K)
annual mean 2000 - 2005
b)
ORCA1 - WOA09-monthly
Δ of Antarctic Sea Surface Salinity (psu)
annual mean 2000 - 2005
c)
ORCA1 - Dong08-ARGO-monthly
Δ of Antarctic Ocean Mixed Layer Thickness Defined by Sigma T (m)
annual mean 2000 - 2005
d)
Δ of Sea Water Temperature for Southern Ocean
ORCA1 - WOA09-monthly
Δ to (K)
ocean depth (m)
date (yyyy-mm)
e)
Δ of Sea Water Salinity for Southern Ocean
ORCA1 - WOA09-monthly
Δ so (psu)
ocean depth (m)
date (yyyy-mm)

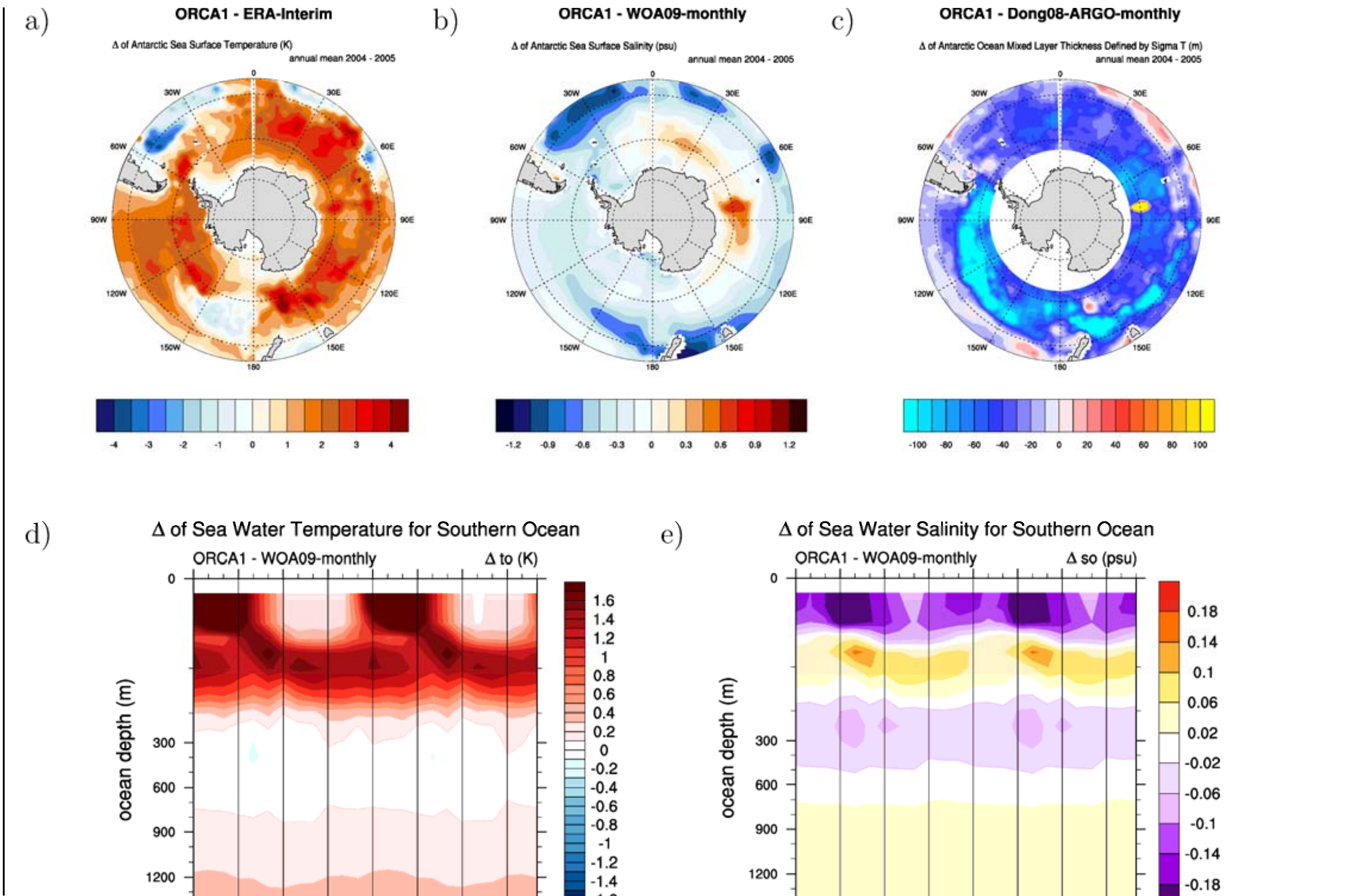

a)
ORCA1 - ERA-Interim
Δ of Antarctic Sea Surface Temperature (K)
annual mean 2004 - 2005
b)
ORCA1 - WOA09-monthly
Δ of Antarctic Sea Surface Salinity (psu)
annual mean 2004 - 2005
c)
ORCA1 - Dong08-ARGO-monthly
Δ of Antarctic Ocean Mixed Layer Thickness Defined by Sigma T (m)
annual mean 2004 - 2005
d)
Δ of Sea Water Temperature for Southern Ocean
ORCA1 - WOA09-monthly
Δ to (K)
ocean depth (m)
e)
Δ of Sea Water Salinity for Southern Ocean
ORCA1 - WOA09-monthly
Δ so (psu)
ocean depth (m)

Figure 14. Annual-mean difference between EC-Earth/NEMO and ERA-Interim sea surface temperatures (a), the World Ocean Atlas sea surface salinity (b), and the Argo float observations for ocean mixed layer thickness (c), showing that in the Southern Ocean SSTs in EC-Earth are too high, sea surface salinity too fresh, and the mixed layer too shallow. The other available diagnostics of the *namelist_SouthernOcean.nml* help ~~in~~ understanding these biases. Vertical sections of temperature (d) and salinity differences (e) reveal that the SST bias is mainly an austral summer problem, but also that vertical mixing is not able to penetrate a year-round existing warm layer below 80 m depth.

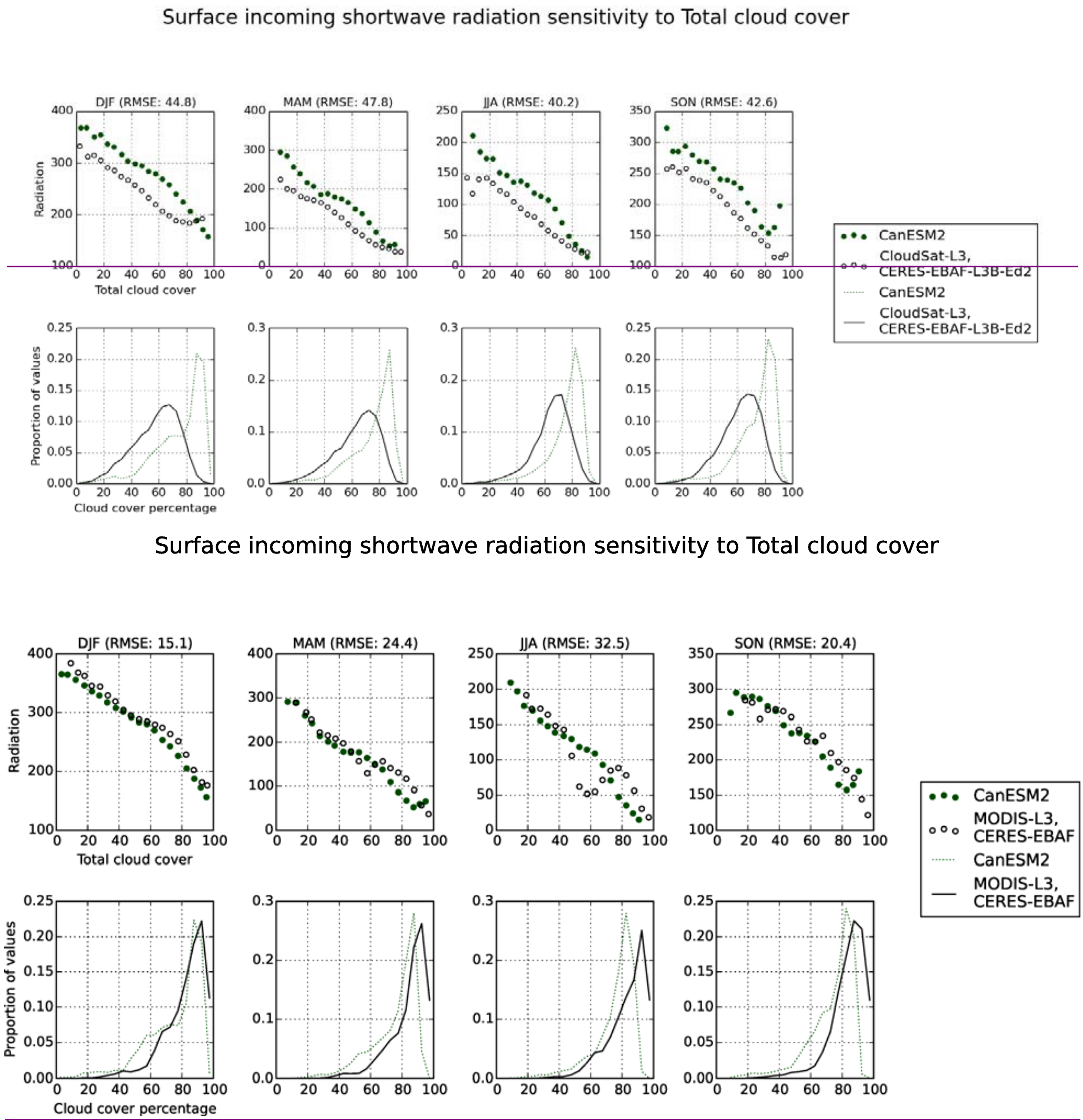
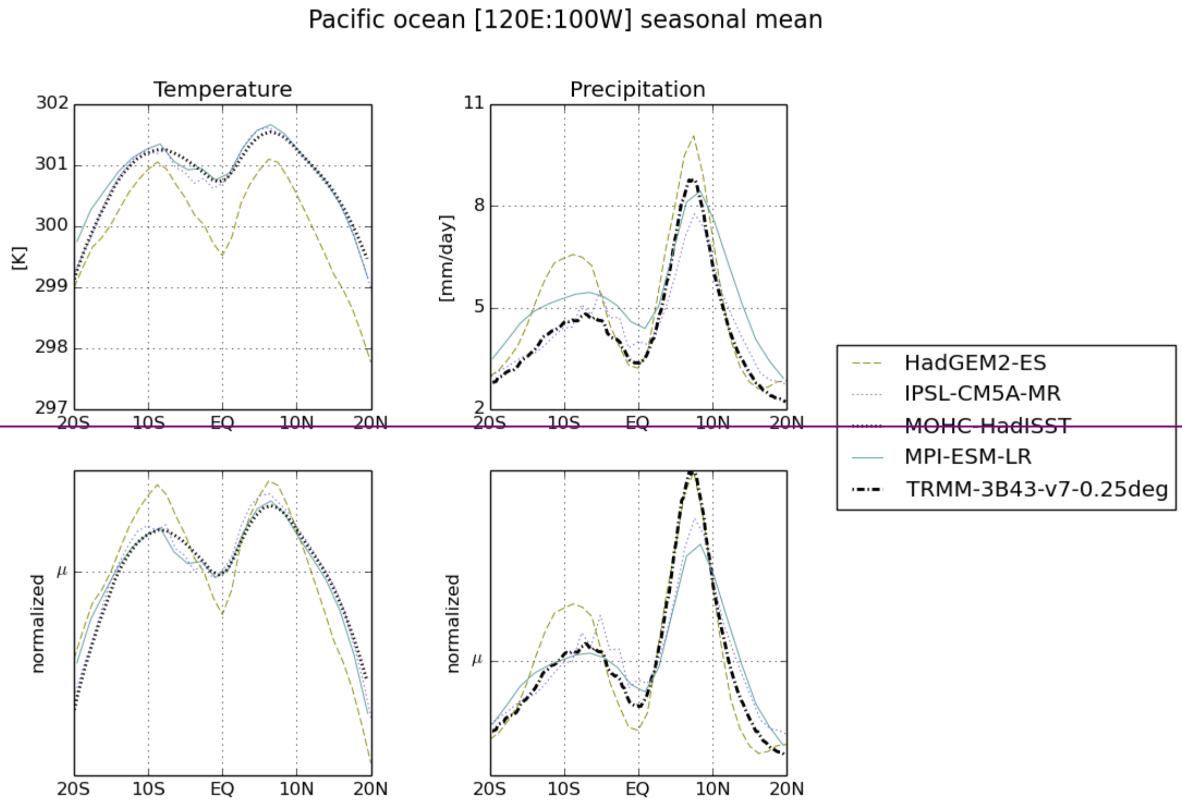


Figure 15. Upper panel: ~~Covariability~~covariability between incoming surface short wave radiation (rsds) and total cloud cover (clt). Lower panel: ~~Fraction~~fraction occurrence histograms of binned cloud cover: ~~Observations~~observations are CERES-EBAF (radiation) and CloudSat (cloud cover). The CanESM2 model from the CMIP5 archive is shown as an example for comparison to observations (the namelists runs on all CMIP5 models). CanESM2 generally reproduces the observed slope of rsds as a function of clt, although there is a systematic positive bias in the amount of shortwave radiation reaching the surface for most cloud cover values. A positive bias is also seen in the CanESM2 histogram of cloud occurrence, with a strong peak in seasonal cloud fraction of

90% in most seasons. Produced with *namelist_~~SouthernAtmosphere~~SouthernHemisphere.xml*.


Pacific ocean [120E:100W] seasonal mean
Temperature
Precipitation
[K]
[mm/day]
normalized
μ
302
301
300
299
298
297
11
8
5
2
20S
10S
EQ
10N
20N
HadGEM2-ES
IPSL-CM5A-MR
MOHC-HadISST
MPI-ESM-LR
TRMM-3B43-v7-0.25deg

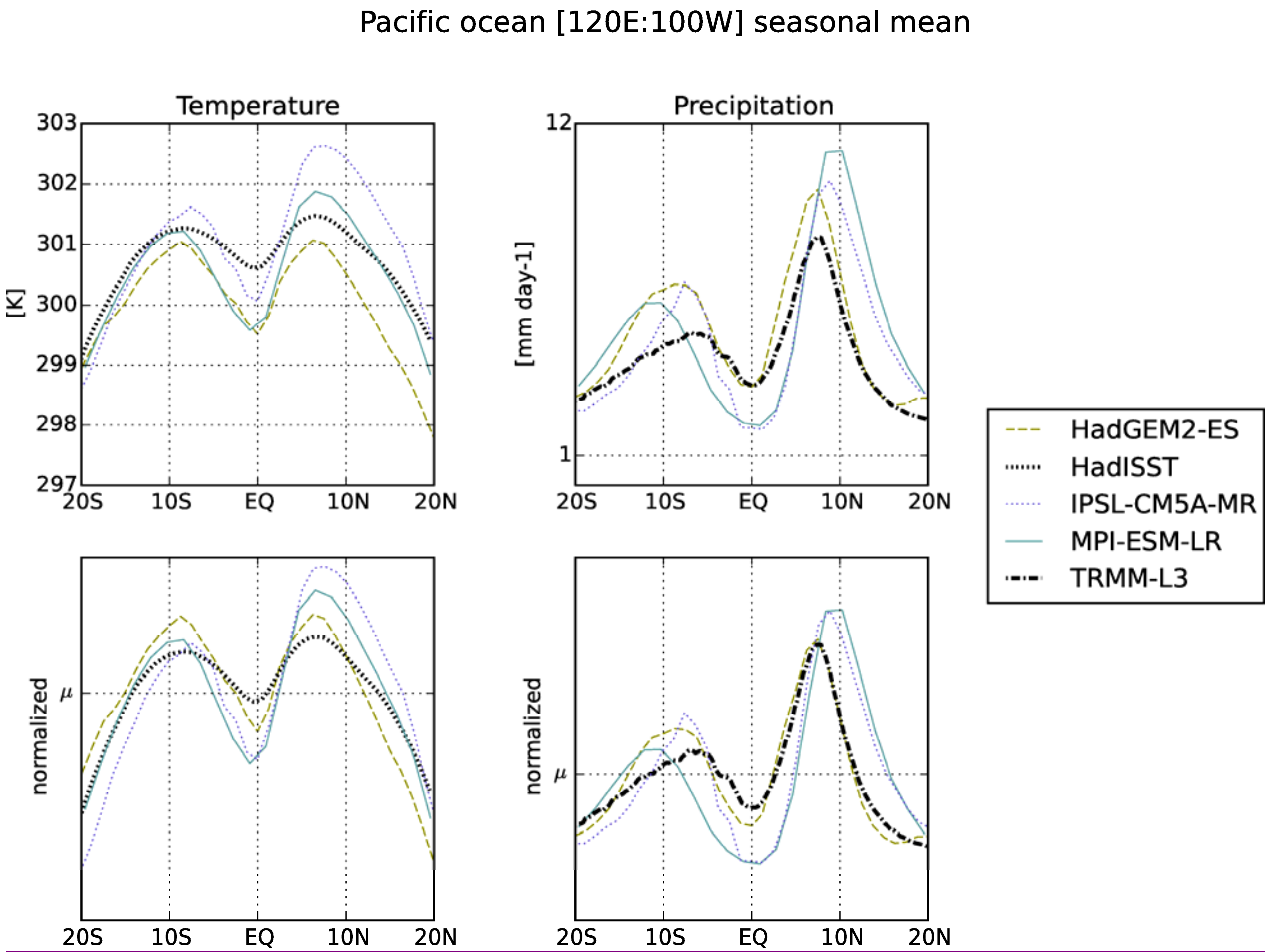
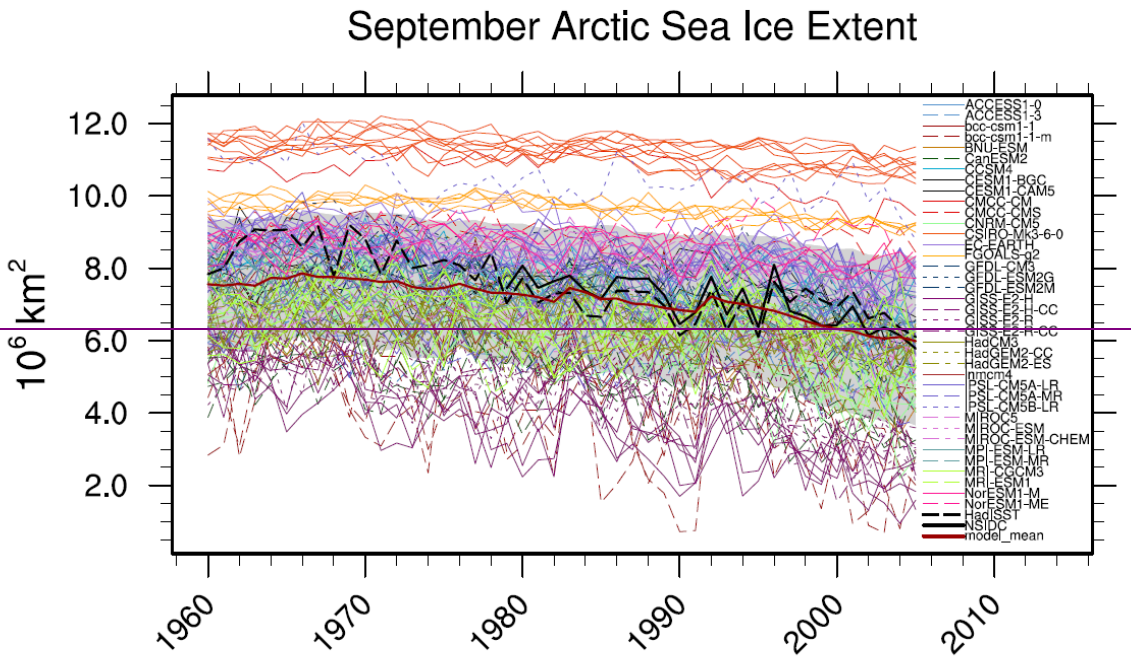


Figure 16. Latitude cross-section of seasonal and zonally averaged values of SSTs and precipitation for the tropical Pacific (zonal averages are made between 120°E and 100°W). Upper panel shows absolute values of SST and precipitation, lower panel shows values normalized by their respective tropical mean value (20°N to 20°~~s~~S) The figure shows that HadGEM2-ES simulates a double ITCZ in the equatorial Pacific with excessive precipitation south of the equator. This bias is accompanied by off equatorial warm biases in normalized SST in both hemispheres and a relative cold bias along the equator. The IPSL-CM5A-MR and MPI-ESM-LR models better capture the SST and precipitation distributions in the tropical Pacific. Produced with *namelist_TropicalVariability.xml*.



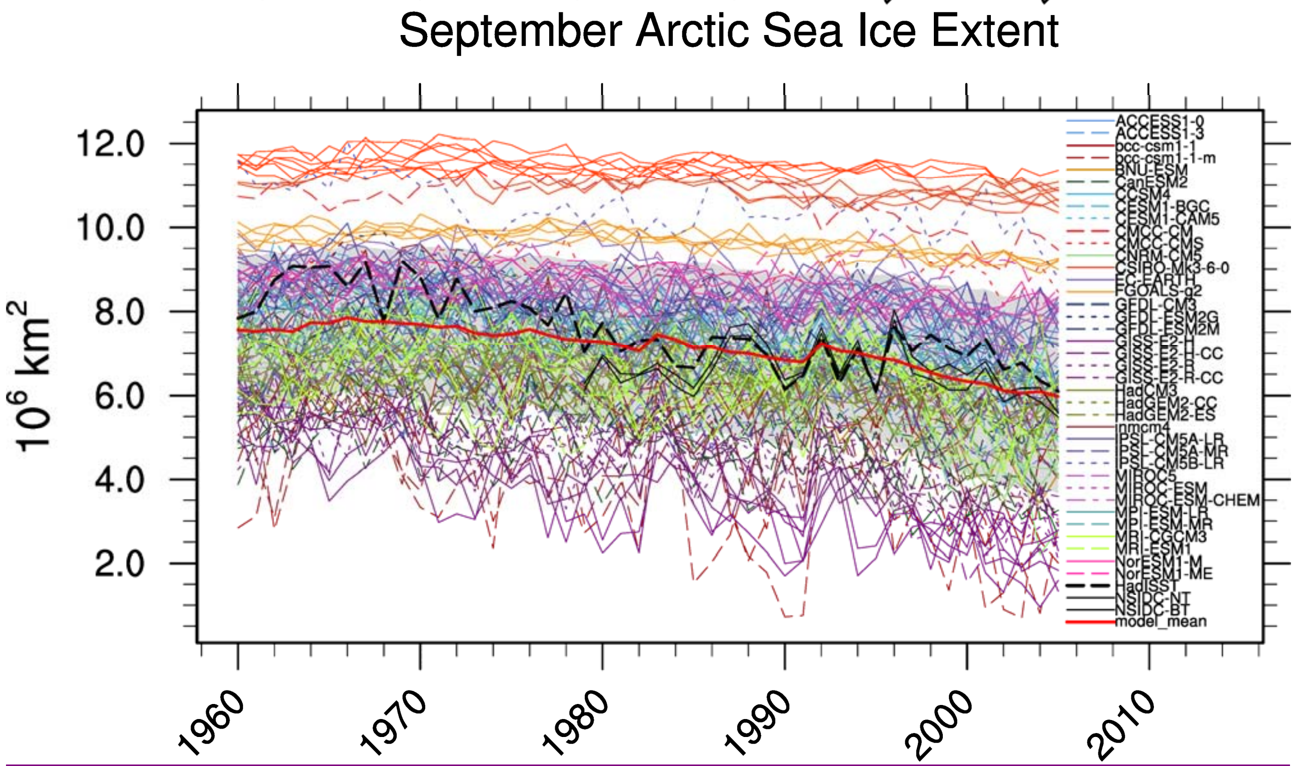
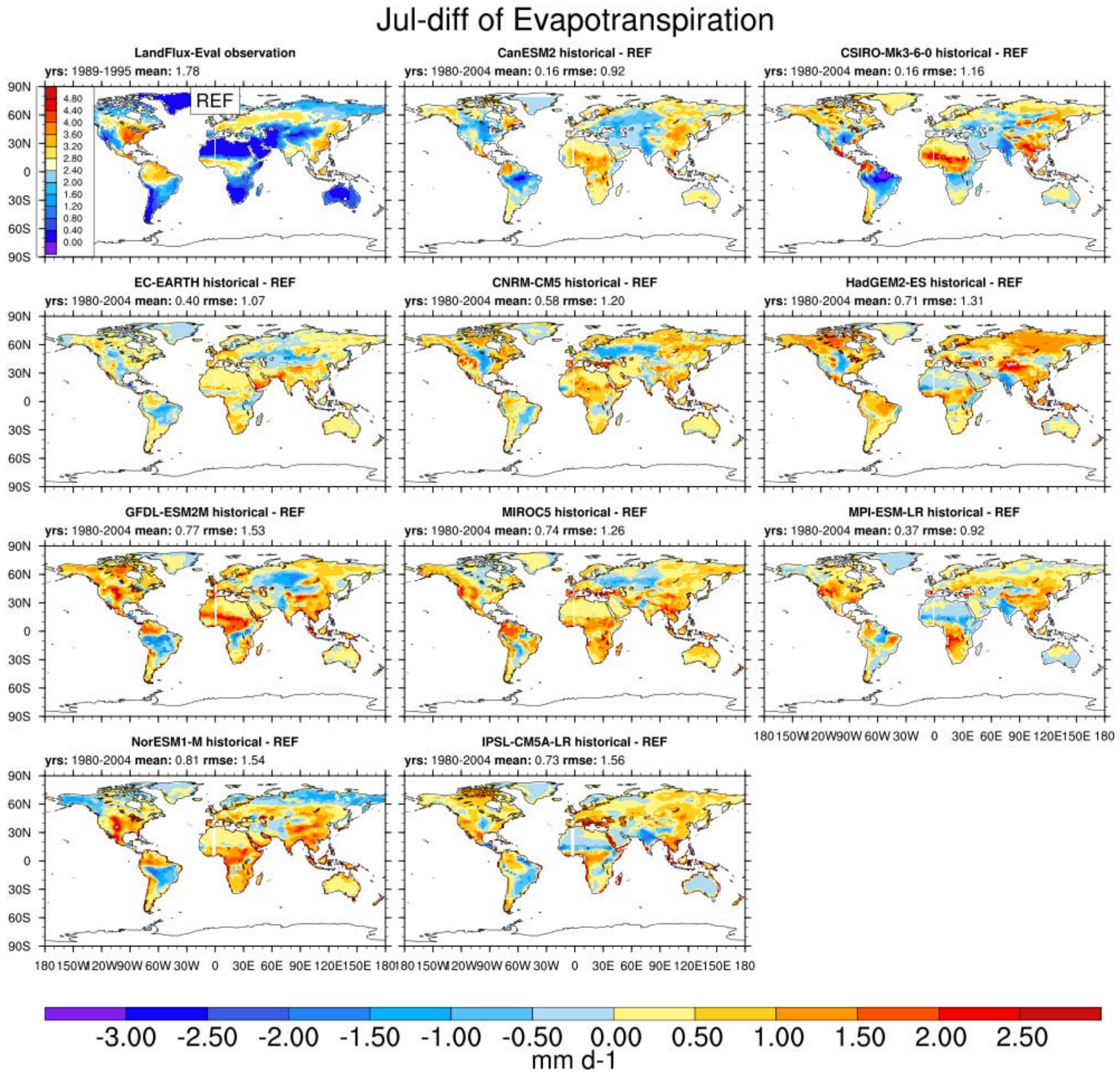


Figure 17. Timeseries (1960-2005) of September mean Arctic sea~~-~~ ice extent from the CMIP5 historical simulations. The CMIP5 ensemble mean is highlighted in dark red and the individual ensemble members of each model (coloured lines) are shown in different linestyles. The model results are compared to observations from the NSIDC (1978-~~2011~~2005, black solid line) and the Hadley Centre Sea ice and Sea Surface Temperature (HadISST, ~~1978-2011~~1960-2005, black dashed line). Consistent with observations, most CMIP5 models show a downward trend in sea ice extent

over the satellite era. The range in simulated sea ice is however quite large (between 3.2 and 12.1 x 10^6 km^2 at the beginning of the timeseries). The multi-model-mean lies below the observations throughout the entire ~~timeseries~~time period, especially after 1978, when satellite observation became available. Similar to upper left panel of Figure 9.24 of (Flato et al. (2013)) and produced with *namelist_SeaIce.nml*.


Jul-diff of Evapotranspiration
LandFlux-Eval observation
yrs: 1989-1995 mean: 1.78
REF
CanESM2 historical - REF
yrs: 1980-2004 mean: 0.16 rmse: 0.92
CSIRO-Mk3-6-0 historical - REF
yrs: 1980-2004 mean: 0.16 rmse: 1.16
EC-EARTH historical - REF
yrs: 1980-2004 mean: 0.40 rmse: 1.07
CNRM-CM5 historical - REF
yrs: 1980-2004 mean: 0.58 rmse: 1.20
HadGEM2-ES historical - REF
yrs: 1980-2004 mean: 0.71 rmse: 1.31
GFDL-ESM2M historical - REF
yrs: 1980-2004 mean: 0.77 rmse: 1.53
MIROC5 historical - REF
yrs: 1980-2004 mean: 0.74 rmse: 1.26
MPI-ESM-LR historical - REF
yrs: 1980-2004 mean: 0.37 rmse: 0.92
NorESM1-M historical - REF
yrs: 1980-2004 mean: 0.81 rmse: 1.54
IPSL-CM5A-LR historical - REF
yrs: 1980-2004 mean: 0.73 rmse: 1.56
-3.00 -2.50 -2.00 -1.50 -1.00 -0.50 0.00 0.50 1.00 1.50 2.00 2.50
mm d-1

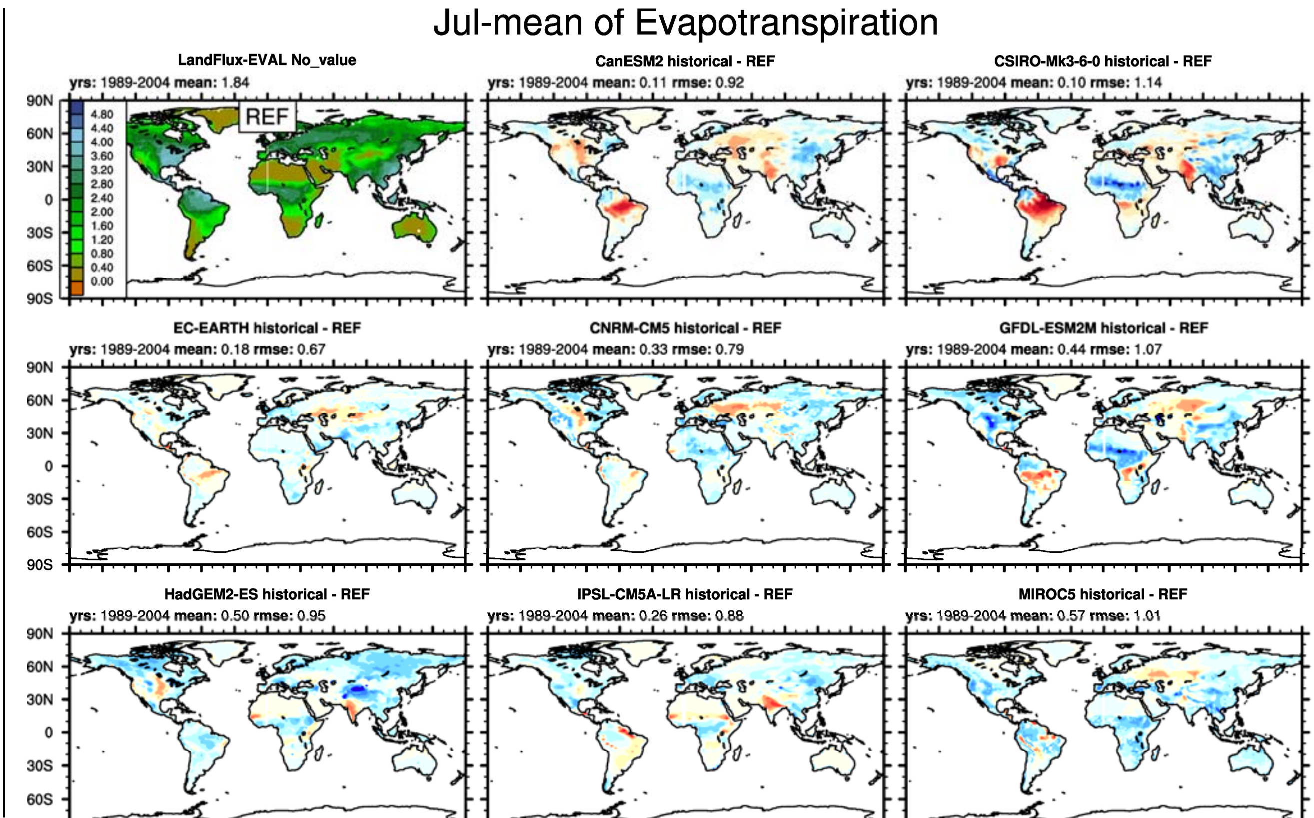
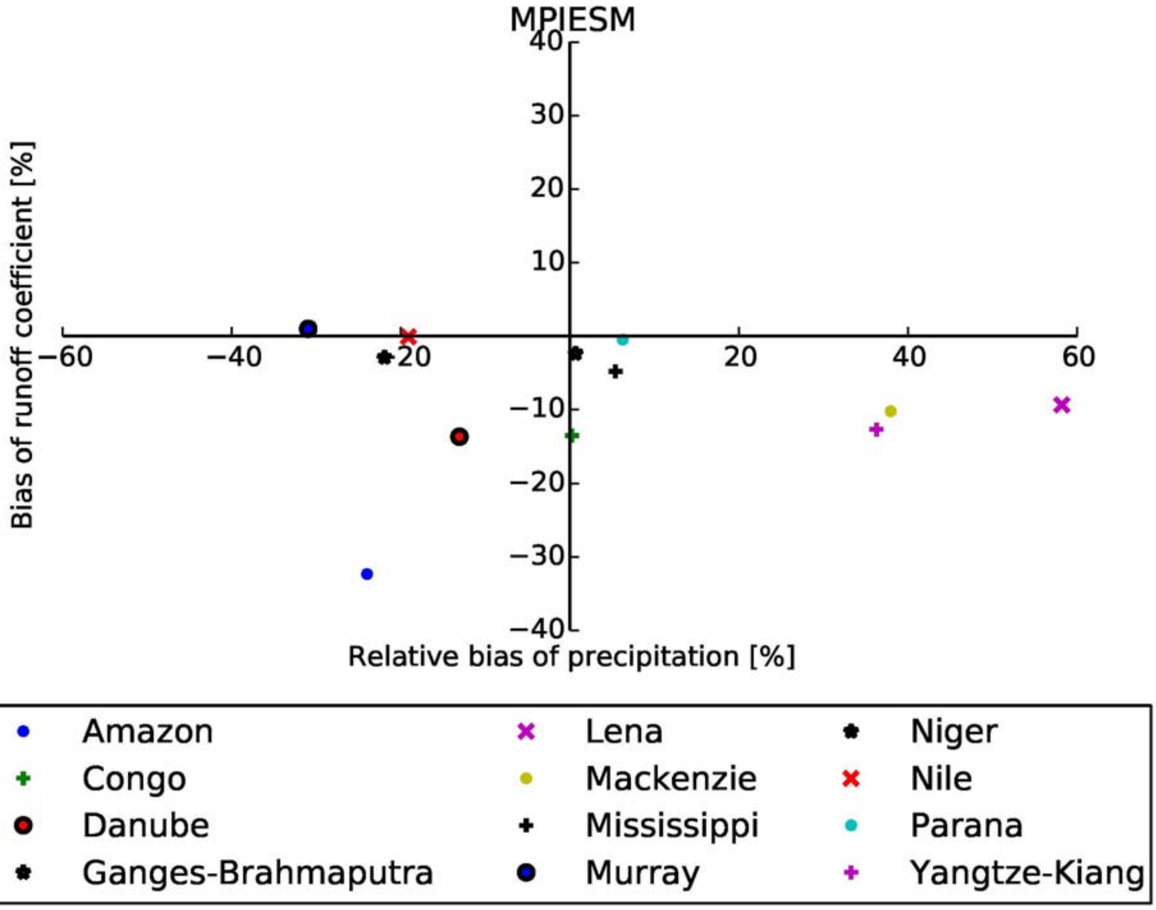

Jul-mean of Evapotranspiration
LandFlux-EVAL No_value
yrs: 1989-2004 mean: 1.84
REF
CanESM2 historical - REF
yrs: 1989-2004 mean: 0.11 rmse: 0.92
CSIRO-Mk3-6-0 historical - REF
yrs: 1989-2004 mean: 0.10 rmse: 1.14
EC-EARTH historical - REF
yrs: 1989-2004 mean: 0.18 rmse: 0.67
CNRM-CM5 historical - REF
yrs: 1989-2004 mean: 0.33 rmse: 0.79
GFDL-ESM2M historical - REF
yrs: 1989-2004 mean: 0.44 rmse: 1.07
HadGEM2-ES historical - REF
yrs: 1989-2004 mean: 0.50 rmse: 0.95
IPSL-CM5A-LR historical - REF
yrs: 1989-2004 mean: 0.26 rmse: 0.88
MIROC5 historical - REF
yrs: 1989-2004 mean: 0.57 rmse: 1.01

Figure 18. Bias in evapotranspiration (mm/~~d~~day) for July in a subset of CMIP5 models in reference to the LandFlux-EVAL evapotranspiration product. The global mean bias is also indicated for each model as well as the RMSE. The comparison reveals the existence of biases in July evapotranspiration for a subset of CMIP5 models. All models overestimate evapotranspiration in summer, especially in Europe, Africa, China, Australia, Western North America, and parts of Amazonia. Biases of the opposite sign (underestimation in evapotranspiration) can be seen in some other regions of the world, notably over parts of the tropics. For most regions, there is a clear correlation between biases in evapotranspiration and precipitation (see precipitation bias in Fig. 4). Produced with *namelist_~~Evapotransport~~Evapotranspiration.xml*.



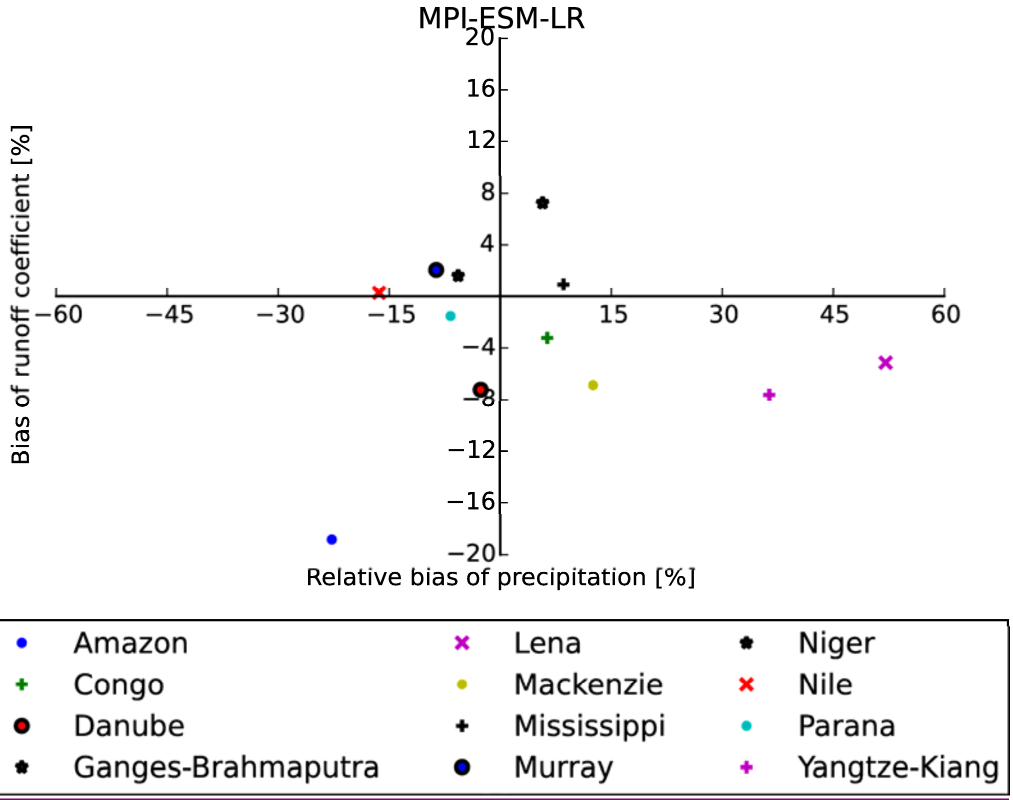
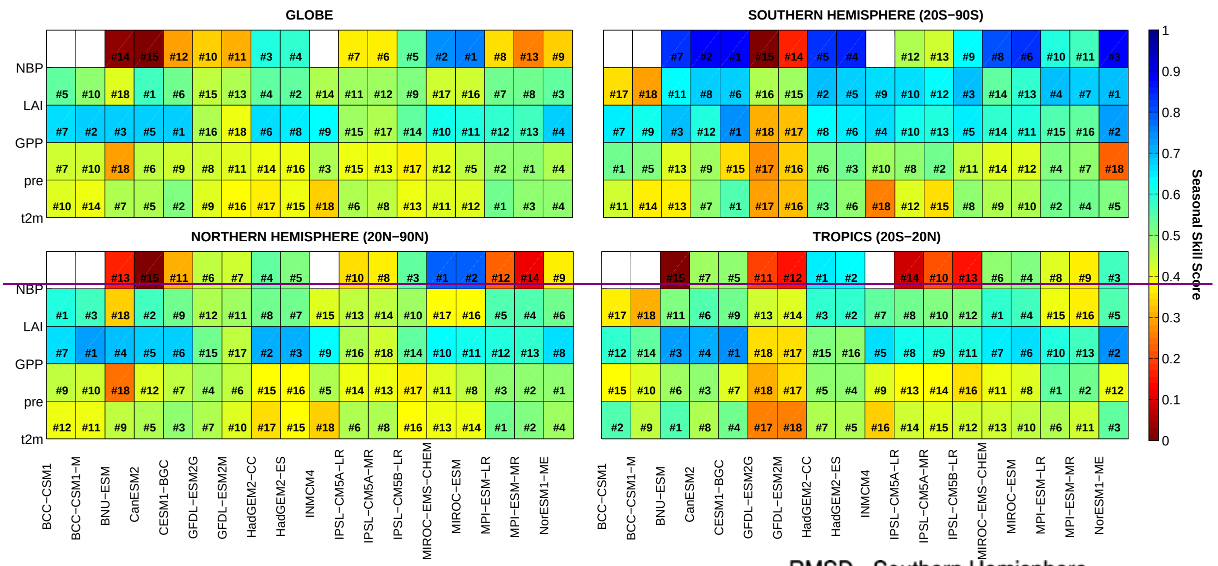


Figure 19. Biases in runoff coefficient (runoff/precipitation) and precipitation for major catchments of the globe. The MPI-ESM~~-1.1~~-LR historical ~~simnulation~~simulation is used as an example. Even though positive and negative precipitation biases exist for MPI-ESM~~-1.1~~-LR in the various catchment areas, the bias in the runoff coefficient is usually negative. This implies that the fraction of evapotranspiration generally tends to be overestimated by the model independently of whether precipitation has a positive or negative bias. Produced with *namelist_runoff_et.xml.*



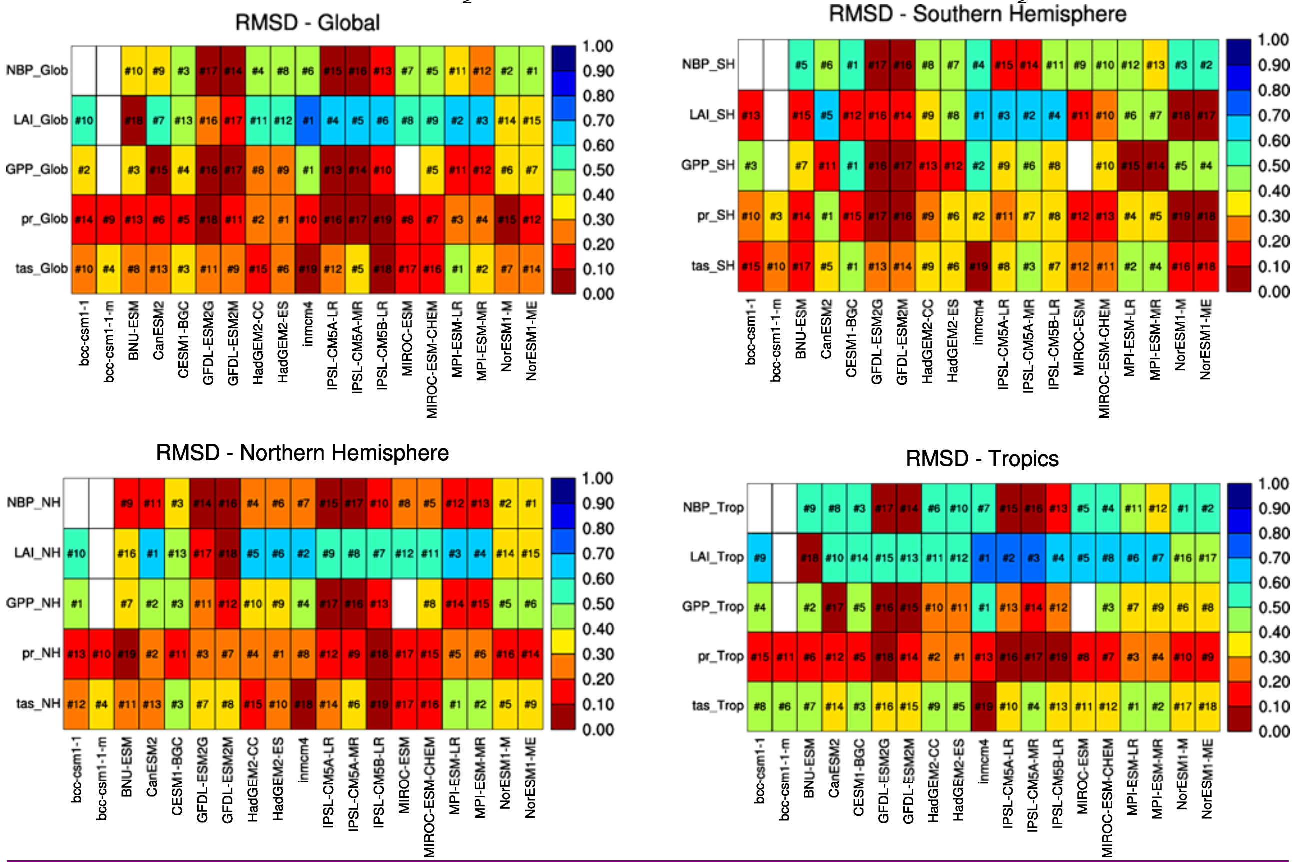
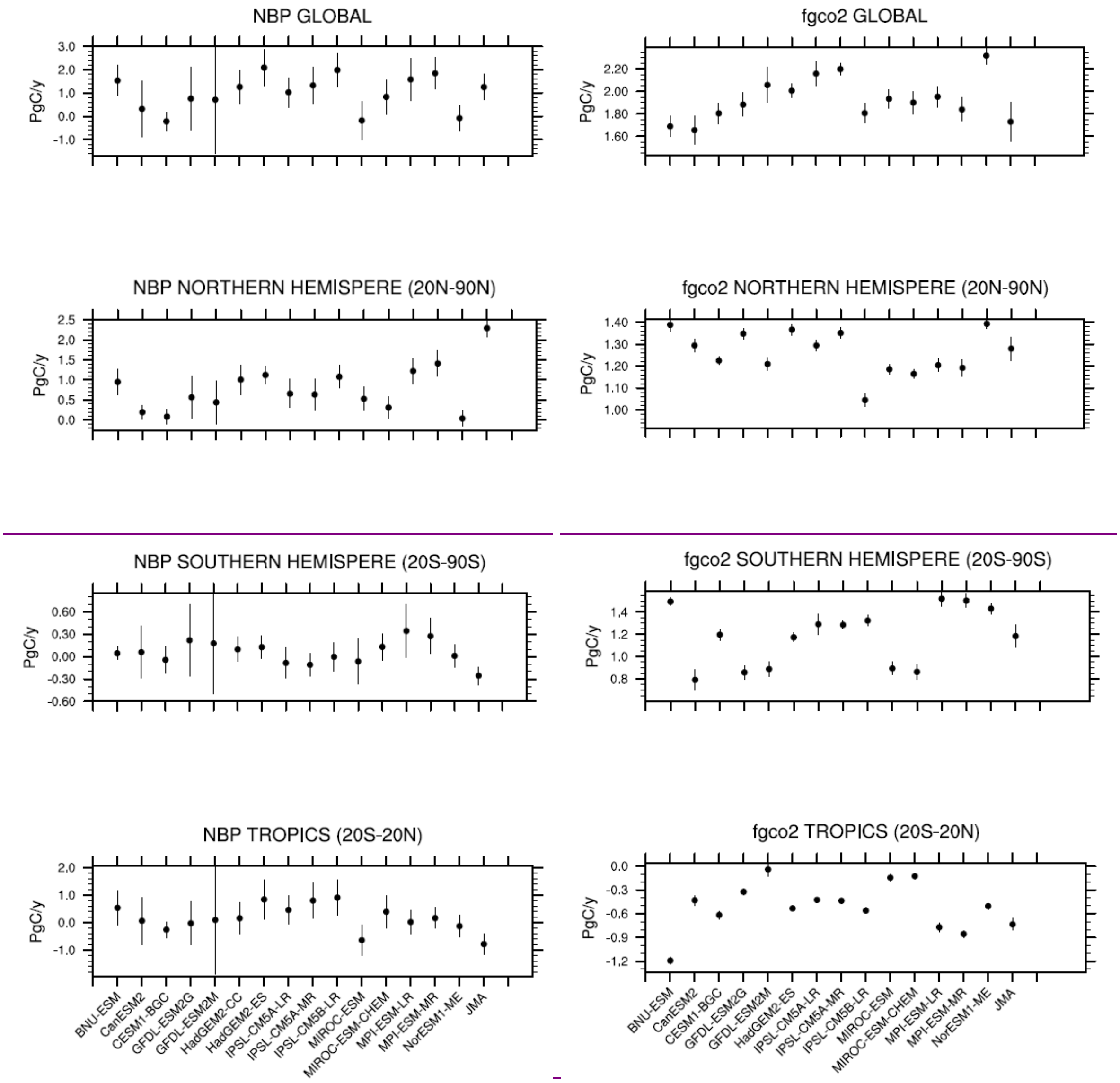


Figure 20. Relative space-time RMSE calculated from the 1986–2005 climatological seasonal cycle of the CMIP5 historical simulations over different sub-domains for NBP, LAI, GPP, precipitation, and near-surface air temperature. The RMSE has been normalized with the maximum RMSE in order to have a skill score ranging between 0 and 1. A score of 0 indicates poor performance of models reproducing the phase and amplitude of the reference mean annual cycle, while a perfect score is equal to 1. The comparison suggests that there is no clearly superior model for all variables.

All models have significant problems in representing some key biogeochemical variables such as NBP and LAI, with largest errors in the tropics mainly because of a too weak seasonality. Similar to Figure 18 of Anav et al. (2013) and ~~reproduced~~produced with *namelist_~~perfmetrics_CMIP5~~anav13jclim.xml*.


NBP GLOBAL
fgco2 GLOBAL
NBP NORTHERN HEMISPERE (20N-90N)
fgco2 NORTHERN HEMISPERE (20N-90N)
NBP SOUTHERN HEMISPERE (20S-90S)
fgco2 SOUTHERN HEMISPERE (20S-90S)
NBP TROPICS (20S-20N)
fgco2 TROPICS (20S-20N)
PgC/y
BNU-ESM
CanESM2
CESM1-BGC
GFDL-ESM2G
GFDL-ESM2M
HadGEM2-CC
HadGEM2-ES
IPSL-CM5A-LR
IPSL-CM5A-MR
IPSL-CM5B-LR
MIROC-ESM
MIROC-ESM-CHEM
MPI-ESM-LR
MPI-ESM-MR
NorESM1-ME
JMA

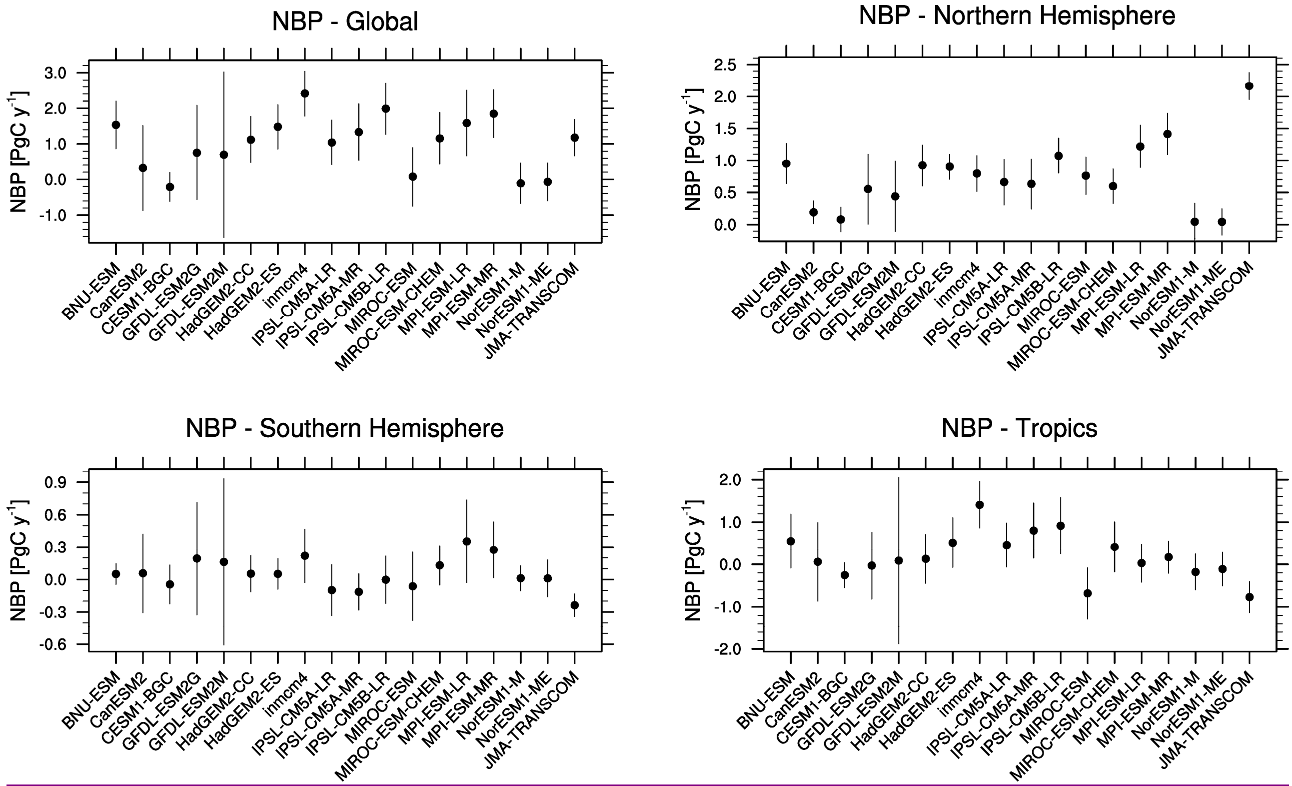
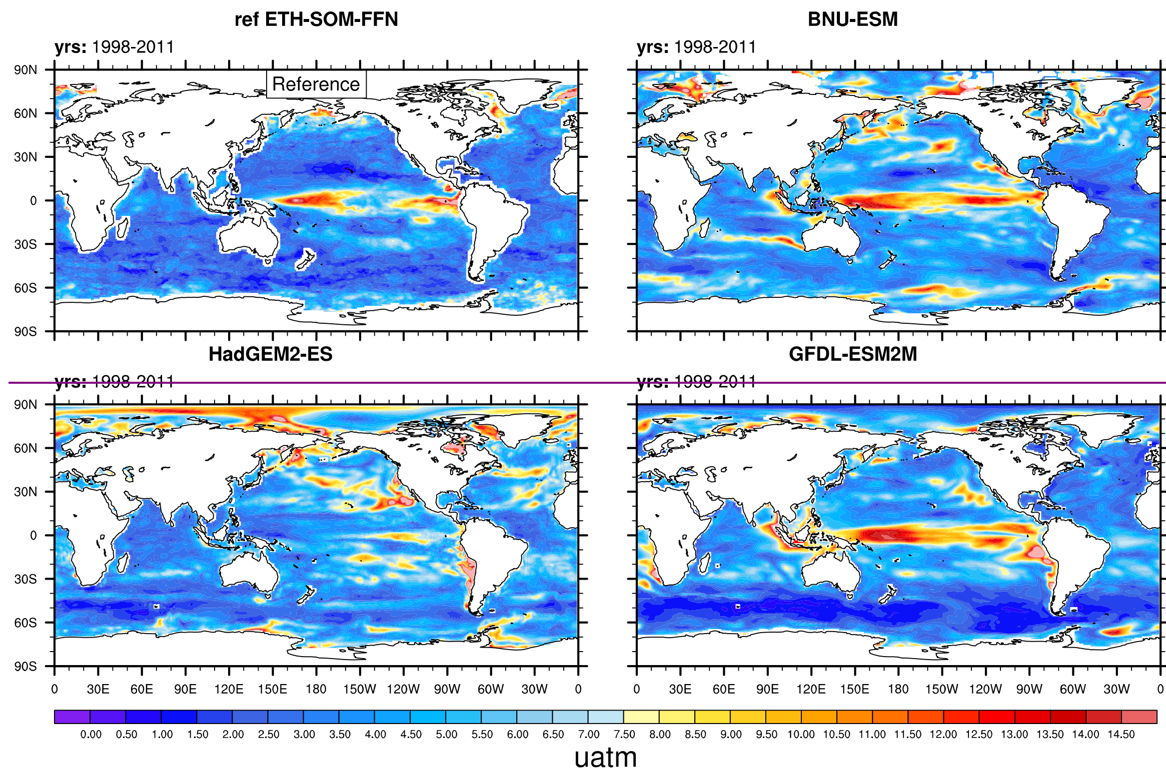


Figure 21. Error-bar plot showing the 1986-2005 CMIP5 integrated NBP ~~(left) and ocean atmosphere carbon fluxes (fgco2, right) over the~~for different land ~~and ocean~~ subdomains~~, respectively.~~. Positive values ~~in~~of NBP ~~and fgco2~~ correspond to land ~~and ocean~~ uptake, ~~respectively, and~~ vertical bars are computed considering the interannual variation. The models are compared to JMA inversion estimates. The models' range is very large and results show that ESMs fail to accurately reproduce the global net land CO_2 flux~~ (NBP, left). In general, ESMs simulate global ocean-atmosphere CO_2 fluxes (fgco2, right) that are comparable to the inversions and GCP estimates.~~. At the hemispheric scale, there is no clear bias common in most ESMs, except in the tropics where models simulate a lower CO_2 source than that estimated by the inversion. Reproducing ~~Figures~~Figure 6 ~~and 14~~ of Anav et al. (2013) and produced with *namelist_anav13jclim.xml.*

Form


ref ETH-SOM-FFN
BNU-ESM
HadGEM2-ES
GFDL-ESM2M
yrs: 1998-2011
Reference
90N
60N
30N
0
30S
60S
90S
0
30E
60E
90E
120E
150E
180
150W
120W
90W
60W
30W
0.00 0.50 1.00 1.50 2.00 2.50 3.00 3.50 4.00 4.50 5.00 5.50 6.00 6.50 7.00 7.50 8.00 8.50 9.00 9.50 10.00 10.50 11.00 11.50 12.00 12.50 13.00 13.50 14.00 14.50
uatm

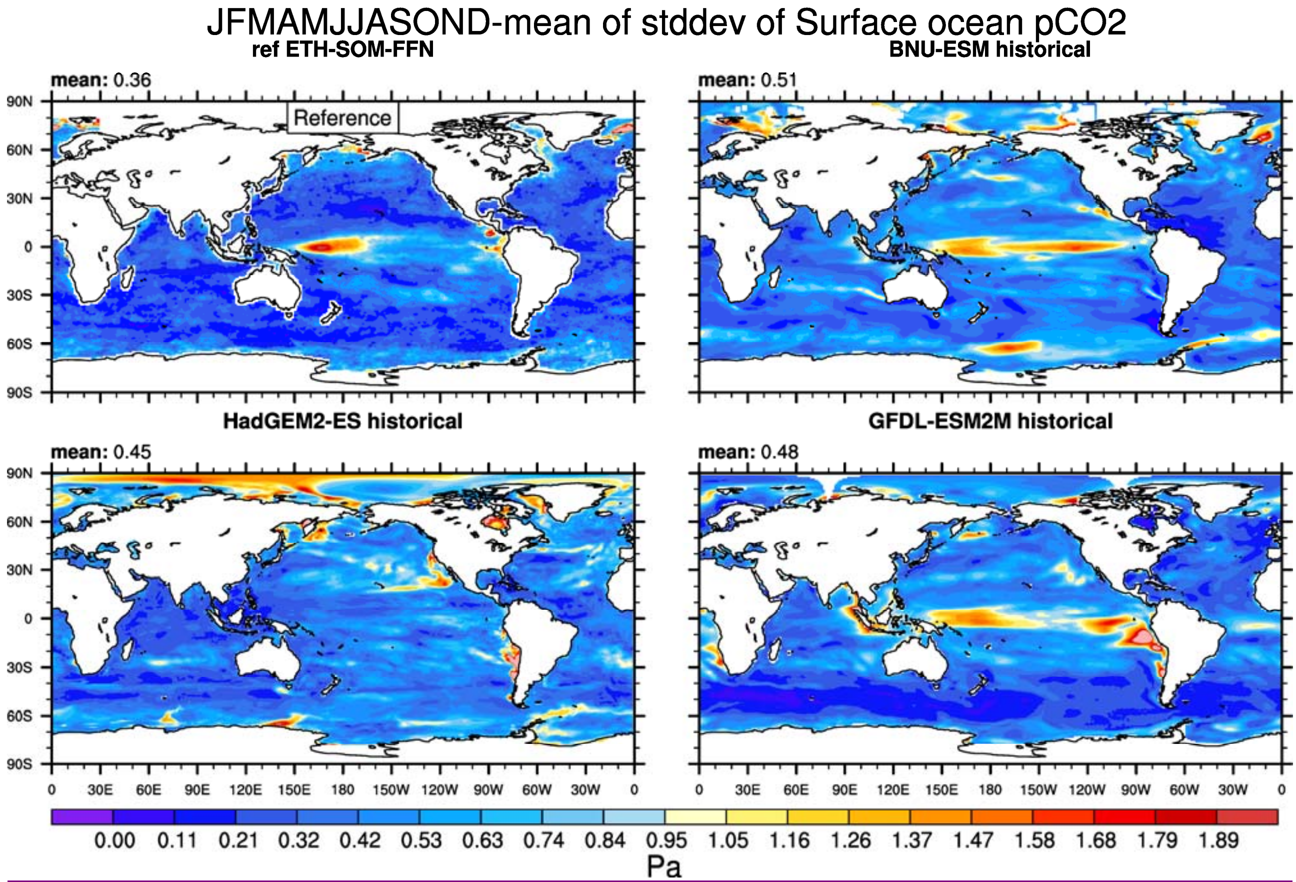
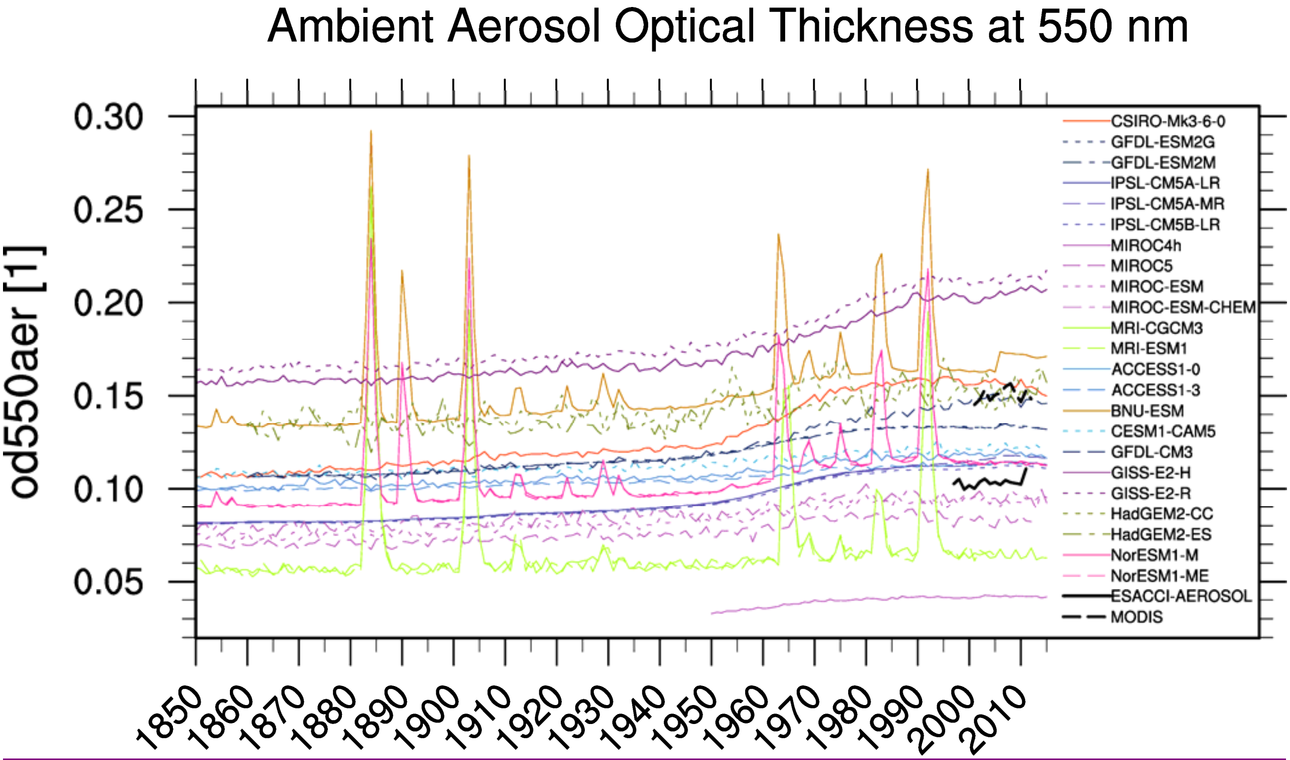

JFMAMJJASOND-mean of stddev of Surface ocean pCO2
ref ETH-SOM-FFN
mean: 0.36
Reference
BNU-ESM historical
mean: 0.51
HadGEM2-ES historical
mean: 0.45
GFDL-ESM2M historical
mean: 0.48
90N
60N
30N
0
30S
60S
90S
0
30E
60E
90E
120E
150E
180
150W
120W
90W
60W
30W
0.00 0.11 0.21 0.32 0.42 0.53 0.63 0.74 0.84 0.95 1.05 1.16 1.26 1.37 1.47 1.58 1.68 1.79 1.89
Pa

Figure 22. Inter-annual variability in de-trended annual mean surface $p\mathrm{CO_2}$ (~~µatm~~*Pa*) for the period 1998~~—~~–2011 from an observation-based reference product (ETH-SOM-FFN; upper left) and three CMIP5 models~~.~~ (1992-2005). The spatial structure of inter-annual variability differs between individual CMIP5 ESMs, however both BNU-ESM and GFDL-ESM2M are able to reproduce pronounced ~~(> 10 µatm)~~ variability in surface ocean *p*CO2 within the Equatorial Pacific, primarily associated with ENSO variability (Rodenbeck et al., 2014). Produced with *namelist_GlobalOcean.xml.*


Ambient Aerosol Optical Thickness at 550 nm
od550aer [1]
0.30
0.25
0.20
0.15
0.10
0.05
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
CSIRO-Mk3-6-0
GFDL-ESM2G
GFDL-ESM2M
IPSL-CM5A-LR
IPSL-CM5A-MR
IPSL-CM5B-LR
MIROC4h
MIROC5
MIROC-ESM
MIROC-ESM-CHEM
MRI-CGCM3
MRI-ESM1
ACCESS1-0
ACCESS1-3
BNU-ESM
CESM1-CAM5
GFDL-CM3
GISS-E2-H
GISS-E2-R
HadGEM2-CC
HadGEM2-ES
NorESM1-M
NorESM1-ME
ESACCI-AEROSOL
MODIS

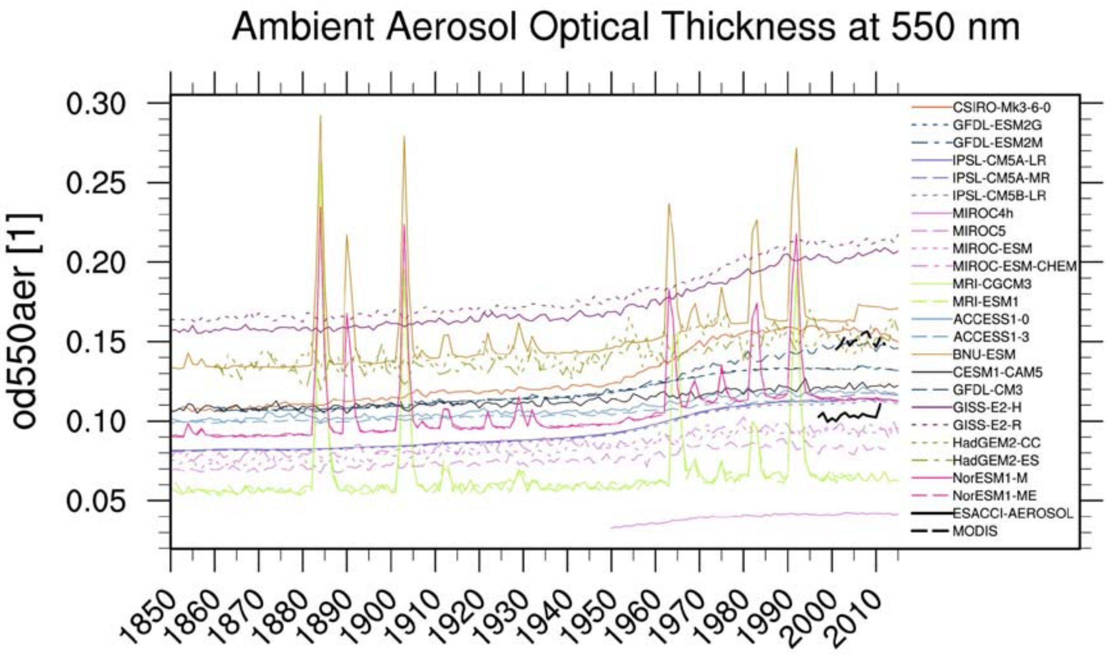


Figure 23. Timeseries of global oceanic mean aerosol optical depth (AOD) from individual CMIP5 models' historical (1850–2005) and RCP 4.5 (2006–2010) simulations, compared with MODIS and ESACCI-AEROSOL satellite data. All models simulate a positive trend in AOD starting around 1950. Some models also show distinct AOD peaks in response to major volcanic eruptions, e.g. El Chichon (1982) and Pinatubo (1991). The models simulate quite a wide range of AODs, between 0.05 and 0.20 in 2010, which largely deviates from the observed values from MODIS and ESACCI-AEROSOL. A significant difference, however, ~~exist~~exists also between the two satellite data ~~set~~sets (about 0.05), indicating an observational uncertainty. Similar to Figure 9.29 of (Flato et al. (2013)) and produced with *namelist_aerosol_CMIP5.xml*.

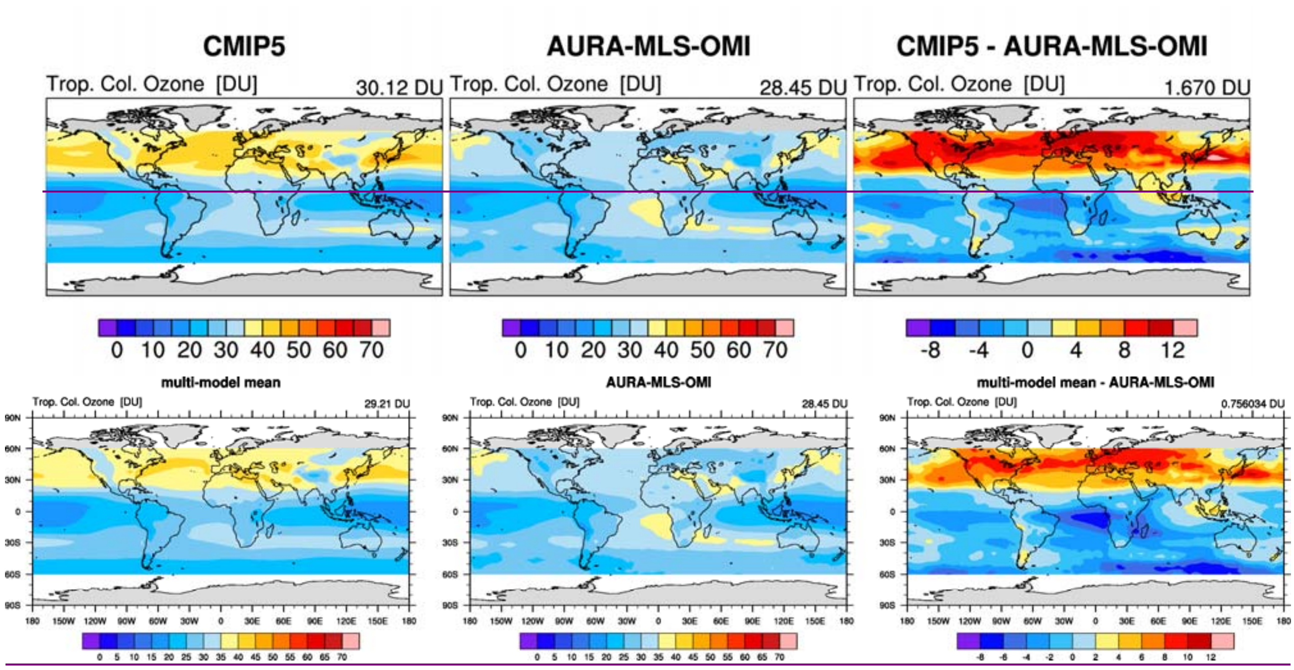
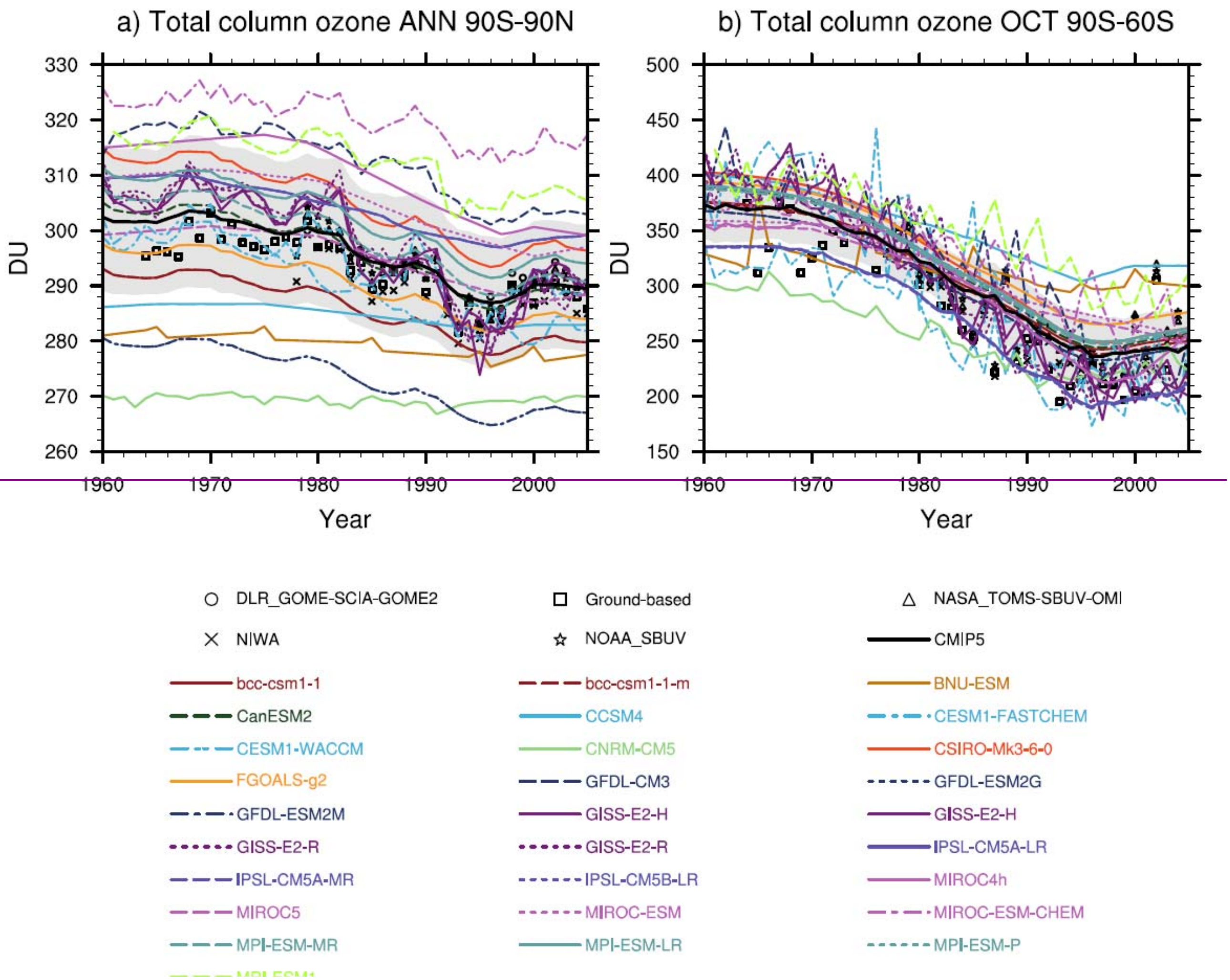


Figure 24. Climatological mean annual mean tropospheric column ozone averaged between 2000 and 2005 from the CMIP5 historical simulations compared to MLS/OMI observations~~.~~ (2005-2012). The values on top of each panel show the global (area-weighted) average, calculated after regridding the data to the horizontal grid of the model and ignoring the grid cells without available observational data. The comparison shows a high bias in tropospheric column ozone in the Northern Hemisphere and a low bias in the Southern Hemisphere in the CMIP5 multi-model mean. Similar to Figure 13 of Righi et al. (2015) and produced with *namelist_righi15gmd_tropo3.xml.*



a) Total column ozone ANN 90S-90N
b) Total column ozone OCT 90S-60S
DU
Year
DLR_GOME-SCIA-GOME2
Ground-based
NASA_TOMS-SBUV-OMI
NIWA
NOAA_SBUV
CMIP5
bcc-csm1-1
bcc-csm1-1-m
BNU-ESM
CanESM2
CCSM4
CESM1-FASTCHEM
CESM1-WACCM
CNRM-CM5
CSIRO-Mk3-6-0
FGOALS-g2
GFDL-CM3
GFDL-ESM2G
GFDL-ESM2M
GISS-E2-H
GISS-E2-H
GISS-E2-R
GISS-E2-R
IPSL-CM5A-LR
IPSL-CM5A-MR
IPSL-CM5B-LR
MIROC4h
MIROC5
MIROC-ESM
MIROC-ESM-CHEM
MPI-ESM-MR
MPI-ESM-LR
MPI-ESM-P

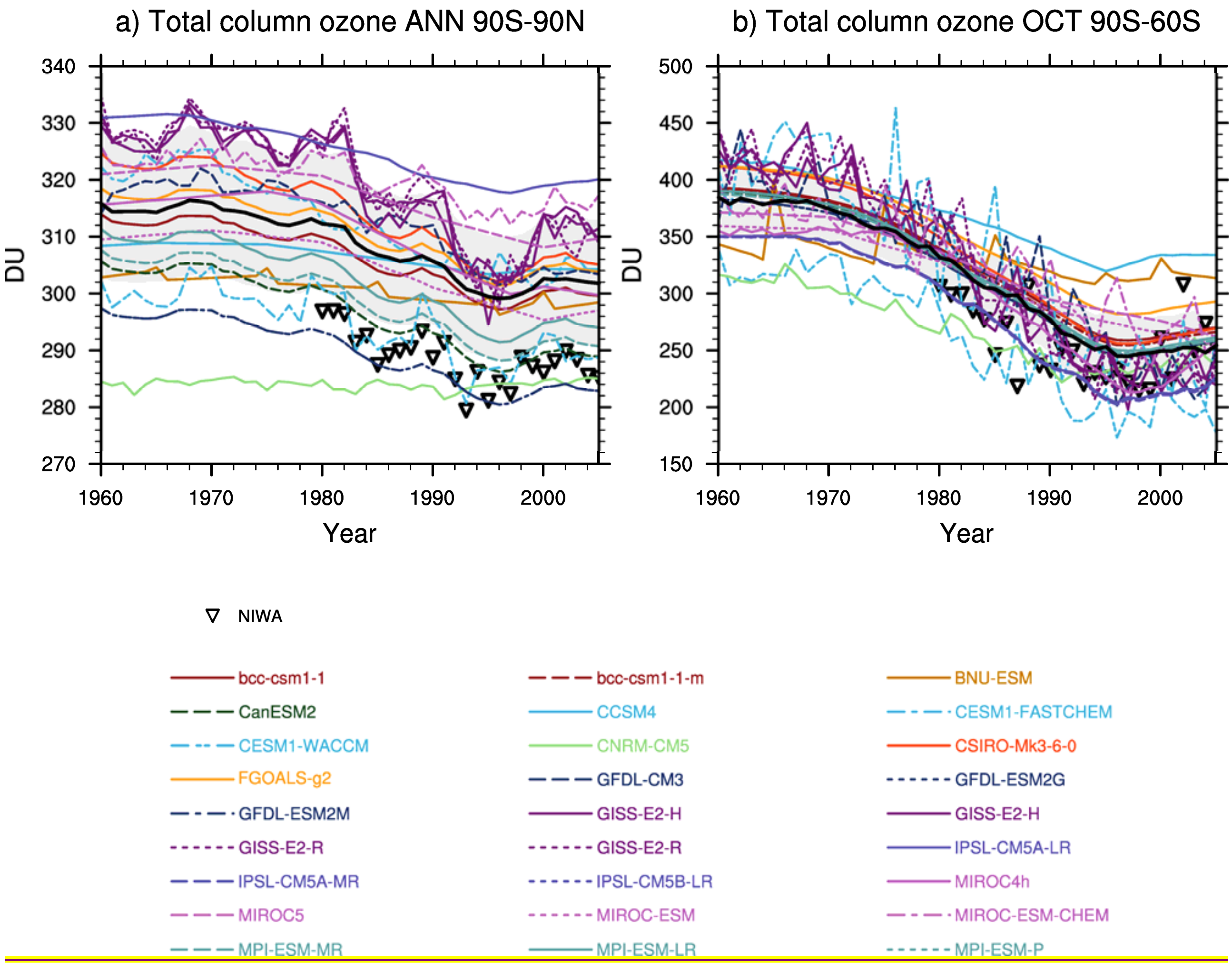
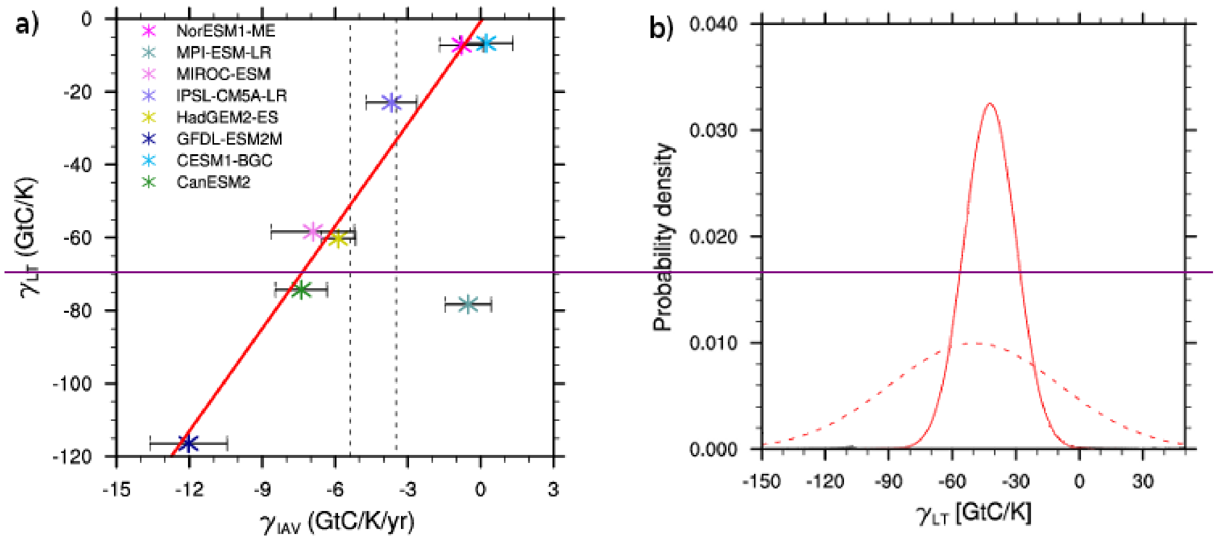


Figure 25. Total column ozone time series for (a) annual global and (b) Antarctic October mean. CMIP5 models are shown in coloured lines and the multi-model mean in thick black, their standard deviation as grey shaded area, and observations from ~~five different sources~~NIWA (black ~~symbols~~triangles). The CMIP5 multi-model mean is in good agreement with observations, but significant deviations exist for individual models with interactive chemistry. ~~Based on Fig.2 of Eyring et al. (2013)~~Based on Figure 2 of Eyring et al. (2013) and reproducing Figure 9.10 of (Flato et al. (2013)), with *namelist_eyring13jgr.xml*.



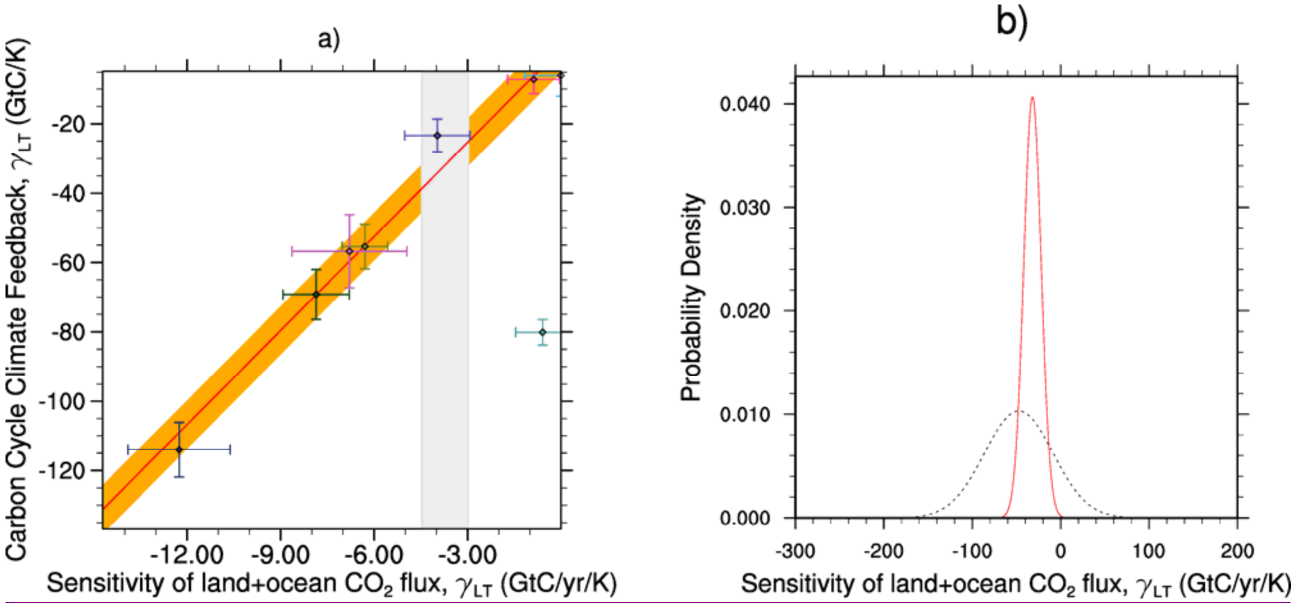


Figure 26. (a) The carbon cycle-climate feedback (γ_{LT}) versus the short-term sensitivity of atmospheric CO_2 to interannual temperature variability (γ_{IAV}) in the tropics for CMIP5 models. The red line shows the best fit line across the CMIP5 simulations and the vertical dashed lines show the observed range of γ_{IAV}. (b) probability distribution function (PDF) for γ_{LT}. The solid line is derived after applying the interannual variability (IAV) constraint to the models while the dashed line is the prior PDF derived purely from the models before applying the IAV constraint. The results show a tight correlation between γ_{LT} and γ_{IAV} that enables the projections to be constrained with observations. The conditional PDF sharpens the range of γ_{LT} to -44 ± 14 GtC/K compared to the unconditional PDF which is (-49 ± 40 GtC/K). ~~Similar to Figure 9.45 of Flato et al. (2013)~~Similar to Figure 9.45 of Flato et al. (2013) and reproducing the CMIP5 model results from Figure 5 of (Wenzel et al. (2014)) with *namelist_wenzel14jgr.xml.*

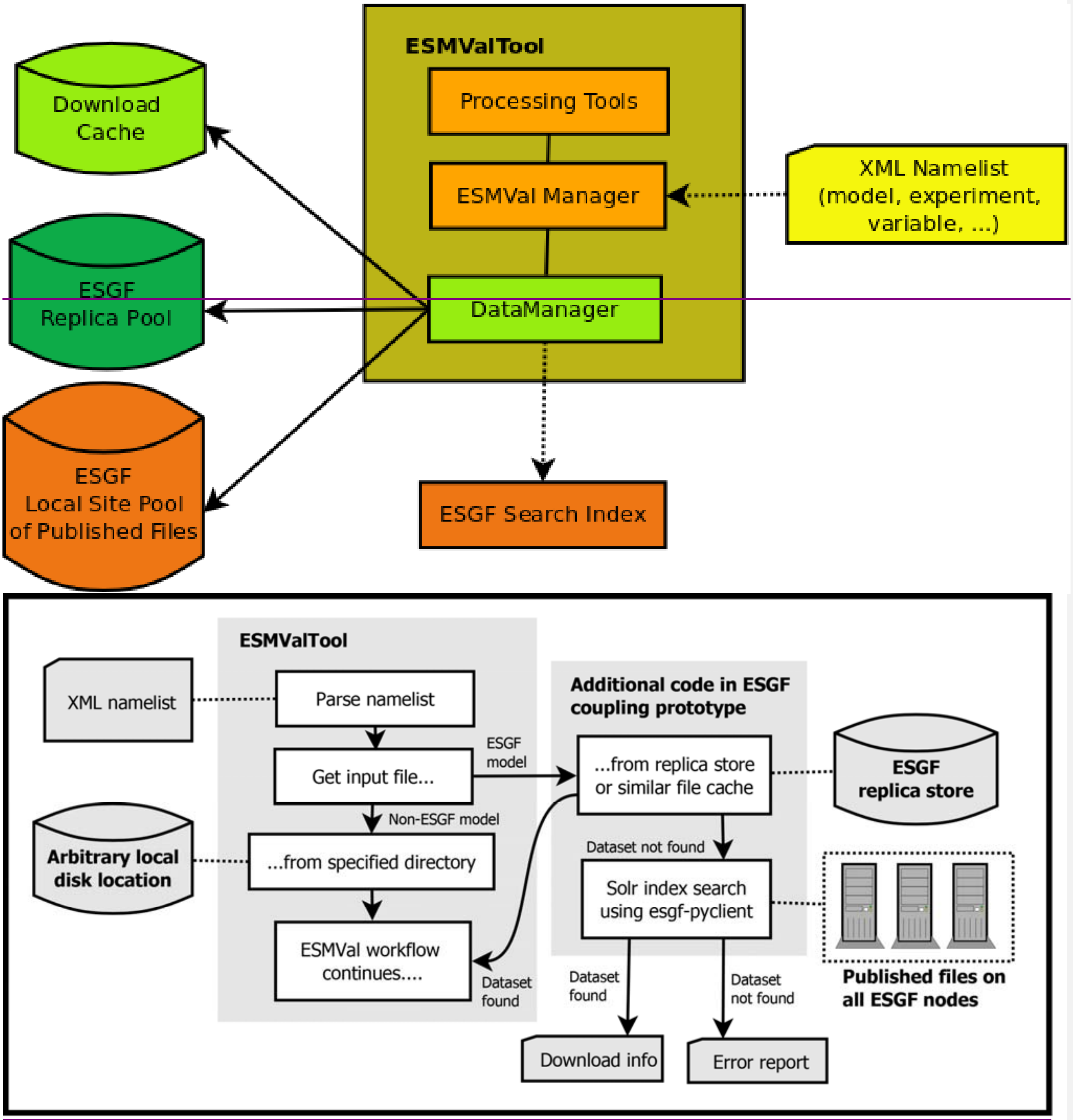


Figure 27. Schematic overview of the coupling of the ESMValTool to the ESGF.