



Earth System Model Evaluation Tool (ESMValTool) v2.0 – diagnostics for extreme events, regional and impact evaluation and analysis of Earth system models in CMIP

Katja Weigel^{1,2}, Lisa Bock², Bettina K. Gier^{1,2}, Axel Lauer², Mattia Righi², Manuel Schlund², Kemisola Adeniyi^{1,2}, Bouwe Andela³, Enrico Arnone^{4,5}, Peter Berg⁶, Louis-Philippe Caron⁷, Irene Cionni⁸, Susanna Corti⁴, Niels Drost³, Alasdair Hunter⁷, Llorenç Lledó⁷, Christian Wilhelm Mohr^{9,10}, Aytac Paçal², Núria Pérez-Zanón⁷, Valeriu Predoi¹¹, Marit Sandstad⁹, Jana Sillmann⁹, Andreas Sterl¹², Javier Vegas-Regidor⁷, Jost von Hardenberg^{13,4}, and Veronika Eyring^{2,1}

¹University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany.

²Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany

10 ³Netherlands eScience Center (NLeSC), Amsterdam, the Netherlands

⁴Institute of Atmospheric Sciences and Climate, Consiglio Nazionale delle Ricerche (ISAC-CNR), Italy

⁵Department of Physics, University of Torino, Italy

⁶Hydrology research unit, Swedish Meteorological and Hydrological Institute (SMHI), Sweden

⁷Barcelona Supercomputing Center (BSC), Barcelona, Spain

15 ⁸Agenzia nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile (ENEA), Rome, Italy

⁹CICERO - Center for International Climate Research, Oslo, Norway

¹⁰Now at: Division for Forestry and Forest Resources, The Norwegian Institute of Bioeconomy Research (NIBIO), Ås, Norway

¹¹NCAS Computational Modelling Services (CMS), University of Reading, Reading, UK

¹²Royal Netherlands Meteorological Institute (KNMI), de Bilt, The Netherlands

20 ¹³Department of Environment, Land and Infrastructure Engineering, Politecnico di Torino, Turin, Italy

Correspondence to: Katja Weigel (weigel@iup.physik.uni-bremen.de)

Abstract. This paper complements a series of now four publications that document the release of the Earth System Model Evaluation Tool (ESMValTool) v2.0. It describes new diagnostics on the hydrological cycle, extreme events, impact assessment, regional evaluations, and ensemble member selection. The diagnostics are developed by a large community of scientists aiming to facilitate the evaluation and comparison of Earth System Models (ESMs) which are participating in the Coupled Model Intercomparison Project (CMIP). The second release of this tool aims to support the evaluation of ESMs participating in CMIP Phase 6 (CMIP6). Furthermore, data sets from other models and observations can be analysed. The diagnostics for the hydrological cycle include several precipitation and drought indices, as well as hydroclimatic intensity and indices from the Expert Team on Climate Change Detection and Indices (ETCCDI). The latter are also used for identification of extreme events and for impact assessment, and to project and characterize the risks and impacts of climate change for natural and socio-economic systems. Further impact assessment diagnostics are included to compute daily temperature ranges and capacity factors for wind and solar energy generation. Regional scales can be analysed with new diagnostics implemented for selected regions and stochastic downscaling. ESMValTool v2.0 also includes diagnostics to analyse large multi-model

25
30



ensembles including grouping and selecting ensemble members by user specified criteria. Here, we present examples for their capabilities based on the well-established CMIP Phase 5 (CMIP5) data set.

1 Introduction

Climate change is affecting the Earth system in many different ways. To be able to assess the impacts of climate change on society and to develop strategies for mitigation and adaptation, a detailed knowledge of the climate system and the key processes driving climate change is necessary. This is particularly the case for changes in the hydrological cycle and climate extreme events, both having direct consequences on ecosystems and society (Eyring et al., 2020). With rising greenhouse gas concentrations the hydroclimatic regime is expected to change (Giorgi et al., 2019). As the intensity and distribution of precipitation determines the availability of fresh water in a certain region, it is also related to the severity of hazardous events such as flooding or droughts. The impact of extreme events on many socio-economic factors increases with their severity, but the rare occurrence of these events makes an assessment of the effect of climate change on such events challenging (Zhang et al., 2011). Especially compound events, caused by a combination of processes on multiple spatial and temporal scales, lead to severe impacts (Zscheischler et al., 2018).

Changes in climate can alter both the strength and the probability of extreme events (Seneviratne et al., 2012; IPCC, 2012). For various extreme events an increase in severity and frequency was observed in the past decades and is expected with rising temperatures, as, for instance, for warm temperature extremes (Alexander, 2016). With rising temperatures an increase is also expected for the amount of precipitation. For wet precipitation extremes this increase is expected to happen faster than for the total wet-day (days with precipitation > 1mm) precipitation (Sillmann et al., 2013b). Several studies project that dry regions become drier and wet regions wetter (Martin, 2018; Greve et al., 2014), which is expected to result in an increase in both wet and dry extreme events, depending on the region. This tendency was highlighted by a general increase of the hydroclimatic intensity, which gives a joint measure of dry and wet conditions in a warming climate (Giorgi et al., 2011). Studies by Donat et al. (2019) and Pfahl et al. (2017) show an increase in observed precipitation extremes in humid regions whereas there is no clear indication on the change of precipitation extreme events in arid regions. The impact of different climate forcings such as greenhouse gases and aerosols on droughts remains to be understood in more detail (Marvel et al., 2019).

Although the climate system is of global extent, its manifestations have regional and local impacts (IPCC, 2014b). Especially for regional climate changes, robust projections require not only an understanding of the underlying physics and internal variability but also a reduction of model biases is essential (Xie et al., 2015). If model biases are corrected without considering the underlying physical processes, however, downscaling of ESM results to regional scales can result in unwanted artefacts (Maraun et al., 2017). Observed changes on the regional scale depend to a large extent on atmospheric dynamics, therefore the signal of climate change is often smaller than the internal variability (Deser et al., 2012) while large differences are found in the modelled future scenarios (Shepherd, 2014). Stochastic downscaling of precipitation can aid in this direction as the fields



65 at regional scale are derived from the spectral properties of the fields at large scale, with an ability in the reproduction of
extremes even over complex orography (Rebora et al. 2006; D’Onofrio et al. 2014; Terzago et al., 2018). Model ensembles
can be used to quantify uncertainties in climate change projections due to internal variability (Xie et al., 2015) and clustering
analysis to inter-compare and group ensemble members based on similar characteristics and select the most representative
ones, going beyond the biases of individual models (Straus et al. 2007).

70 The Earth System Model Evaluation Tool (ESMValTool) version 2.0 (v2.0) includes diagnostics and performance metrics for
the analysis and evaluation of ESMs with observations. It is developed by a large community, which involves more than 150
scientists from over 60 institutions. Figures and other output produced by the tool include full provenance information to allow
for traceability and reproducibility of the results. The main focus is on the analysis of ESM simulations from the Coupled
Model Intercomparison Project (CMIP) of the World Climate Research Programme (WCRP). CMIP started in 1995 (Meehl et
75 al., 2000) with the aim of providing scientists with comparable coupled model runs based on standardized boundary conditions
(Covey et al. 2003). CMIP results from Phase 5 (CMIP5) (Taylor et al., 2012) are the basis for many assessments in the IPCC’s
Fifth Assessment Report (AR5) (IPCC 2013). Now, data from Phase 6 (CMIP6) (Eyring et al., 2016) are available. With every
phase of CMIP the volume of data increases: for CMIP6 a total data volume of about 20 to 40 PB is expected. This emphasizes
the need for a fast and comprehensive tool like the ESMValTool (v2.0) to evaluate these model results. In this work, the
80 diagnostics which focus on climate impacts are described and their output using the well-established CMIP5 data is shown.
In this study we present diagnostics included in the ESMValTool specifically for the analysis of the hydrological cycle, extreme
events, climate impacts, multi-model ensemble member sub-selection, and regional model evaluation. This article completes
a series of publications documenting ESMValTool v2.0: Righi et al. (2020) describes the technical aspects, (Eyring et al.,
2020) the new large-scale diagnostics, and (Lauer et al., 2020) emergent constraints and diagnostics for future projections from
85 ESMs in CMIP.

This paper is organized as follows: Section 2 describes the model and observation data used. Section 3 presents the
ESMValTool recipes for the analyses of hydroclimatic intensity, droughts, extreme events, model impact evaluation, multi-
model ensemble member sub-selection, and regional model evaluation. It also describes use of the ESMValTool as a post-
processing tool for further downscaling applications. Section 4 closes with a summary.

90 **2 Models and observations**

ESMValTool v2.0 was developed particularly for the analysis of CMIP data (Righi et al., 2020). This work mainly presents
results based on the well-established CMIP5 model ensemble, but other model output and observational data, e.g. provided by
observations for Model Intercomparison Project (obs4MIPs; Teixeira et al., 2014; Waliser et al., 2020), can also be analysed.
As in version v1.0 (Eyring et al., 2016) ESMValTool v2.0 expects input data to be in a Climate and Forecast (CF) Metadata
95 compliant Network Common Data Format (netCDF) following the Climate Model Output Rewrite (CMOR) standard. The



detailed requirements for CMOR can be found in these tables (<http://pcmdi.github.io/cmor-site/tables.html>). Observations used in the evaluation are described in the following sections of the manuscript but should be seen as examples as they can be easily be replaced by other observational datasets given that they also follow the CMOR convention. Reformatting scripts with downloading instructions are provided with the ESMValTool v2.0 to convert many observational datasets to the CMOR standard.

3 Overview of recipes included in ESMValTool v2.0

This section describes the new and extended ESMValTool v2.0 recipes for analysis of extreme events, regional model output, and for applying ESM output in the assessments of the impact of climate change as well as to carry out model ensemble sub-selection. In ESMValTool v2.0, a recipe is a *.yaml file used to define the diagnostics and performance metrics to apply to the simulation output, as well as the datasets and variables used. The ESMValTool is started from the command line, using for example:

```
esmvaltool run esmvaltool/recipes/examples/recipe_python.yaml
```

Where “esmvaltool/recipes/examples/recipe_python.yaml” is one possible recipe. Instead of this example, any other recipe provided with the ESMValTool or created by the user can be used. For a more detailed instructions how to run the tool and modify or create recipes, see the documentation at <https://docs.esmvaltool.org/>.

In the following, the recipes are briefly described and illustrated with example figures using CMIP5 data. All recipes presented in this work are summarized in Table 1, which includes a short description, together with the analysed variables used, the applied diagnostics and their purpose, as well as the references the diagnostics are based on. Because the online documentation for the ESMValTool v2.0 at <https://docs.esmvaltool.org/> was written simultaneously to this paper by the same authors, there is a considerable overlap to this not peer-reviewed document.

Section 3.1 describes recipes for the hydrological cycle, including indices for hydroclimatic intensity and drought detection. In section 3.2 recipes for other extreme events are presented. Recipes for model impact assessment are described in Section 3.3, and recipes for regional model evaluation in Section 3.4. Section 3.5 presents a recipe for the sub-selection of multi-model ensemble members.

3.1 Hydrological cycle

3.1.1 Hydroclimatic intensity and related indices

The Earth’s hydrological cycle is a key element of the climate system with important impacts on the society. For example, the intensity and distribution of precipitation determines the abundance or scarcity of fresh water in a certain region. It is also related to the severity of hazardous events such as flooding or droughts. Several studies have shown an acceleration of the



hydrological cycle and an intensification of both dry and wet extremes in a warming climate (IPCC, 2013). A simple investigation of total precipitation-related quantities can hide some of the most relevant aspects of the hydrological cycle and its extremes, which can be highlighted through the joint use of the concept of hydroclimatic intensity and related indices (e.g., Giorgi et al., 2014). The hydroclimatic intensity (Giorgi et al., 2011), derived as the product of mean daily precipitation and dry spell length normalized over a reference period, offers a joint view of both dry and wet conditions, allowing to uniquely quantify the response in the intensity of the hydrological cycle in a changing climate. The *hyint* (*hydroclimatic intensity*) diagnostic was developed to calculate several indices for hydroclimatic and climate extremes and allow a multi-index evaluation of climate models.

The *recipe_hyint.yml* calculates six indices for evaluating the to global warming response of the hydrological cycle including both, wet and dry extremes. The indices are selected according to Giorgi et al. (2014), including the simple precipitation intensity index (SDII), the maximum dry spell length (DSL) and wet spell length (WSL), the hydroclimatic intensity index (HY-INT, calculated as normalized DSL times normalized SDII), which is a measure of the overall behaviour of the hydroclimatic cycle (Giorgi et al., 2011), and the precipitation area (PA), i.e. the area over which precipitation occurs at any given day (Giorgi et al., 2014). The *recipe_hyint_extreme_events.yml* can also ingest the 27 temperature and precipitation-based Expert Team on Climate Change Detection and Indices (ETCCDI) (Zhang et al., 2011) calculated by the *recipe_extreme_events.yml* to produce a multi-index analysis (see Sec. 3.2 for further details). The diagnostics perform a subsequent analysis calculating time series and trends of the selected indices for predefined continental areas, normalized to a reference period. The linear model (lm) function of R is used to calculate trends. Statistical significance is tested based on a Student's t-test under a non-null coefficients hypothesis. Trend coefficients and their statistics including standard error, p-value and precipitation above the 95% percentile of the reference distribution are stored. The recipe created several plots, amongst others global and regional maps, timeseries with spread, trend lines and summary plots of trend coefficients. Results are stored in netCDF files including relevant information such as normalization functions and thresholds, and as figures. Figures 1 and 2 show examples of an analysis performed with the *hyint* diagnostic. A map of the HY-INT index (Fig. 1) calculated from EC-EARTH model data shows the projected average HY-INT compared to the reference period (1976-2005): hydroclimatic intensity is projected to greatly increase in some regions (e.g., eastern South America, northern Africa and the Arabian peninsula) and to decrease over other regions (e.g., Antarctica, Greenland, central and North-East Asia, central Africa and western and northern South America) with large areas showing only moderate changes. Trends shown in Fig. 2 exhibit a relatively low inter-model spread for HY-INT. The projected increase in HY-INT seen for all models with values ranging around 10% per century models (also reflected as large geographical patterns) can also be seen in the precipitation intensity (SDI) and heavy precipitation indices (R95), the latter with an increase spread between 10 and 30% per century. Precipitation area (PA) is projected to increase by most models, whereas for projected changes in the dry spell length (DSL) and especially in the wet spell length (WSL) models do not agree on the sign of the projected changes, reflected also in high geographical variability (not shown).



3.1.2 Droughts

Three main types of droughts can be separated: (i) meteorological, (ii) hydrological, and (iii) agricultural droughts. Any type of drought needs to be defined in the context of local and seasonal characteristics implying that a drought should be identified rather as anomalous condition than based on an absolute threshold.

Meteorological droughts are a negative anomaly in precipitation. Depending on the local characteristics, a drought can be defined as an extended period of daily precipitation amounts below a given threshold. The threshold value is defined as the minimum amount of precipitation that is needed to recharge the soil moisture content. This approach requires good knowledge of the local and seasonal characteristics of the soil moisture content. However, it is a useful analysis to investigate climate models' distributions of wet/dry periods which are indicative of how well suited the model is to couple to hydrological impact models. E.g., CMIP5 models have been shown to generally underestimate the number of consecutive dry days (Sillmann et al., 2013b; Cheng et al., 2016). The standardized precipitation index (SPI, McKee et al., 1993) describes local precipitation anomalies and is often used to identify Meteorological droughts. The SPI was developed as a replacement of the commonly used Palmer drought indices (Palmer, 1965) to better capture dry and wet anomalies. The SPI is calculated using monthly mean precipitation. Therefore, it does not account for the intensity of single precipitation events and the runoff process. Furthermore, SPI does not account for evaporation from the surface. This implies, that one component of the water fluxes at the surface is lacking, which makes the SPI incompatible with the concept of hydrological droughts. Evaluation of SPI from CMIP5 models shows large model biases (Ukkola et al., 2018).

A hydrological drought occurs when low water supply effects streams, reservoirs, and groundwater levels and is usually caused by extended periods of meteorological droughts. These hydrological processes are usually not simulated with sufficient details in climate models. As a consequence, also agricultural droughts (i.e. when crops become affected by the hydrological drought) cannot be simulated properly by the models. Hydrological droughts can, however, be estimated in climate models by accounting for evapotranspiration. This allows to estimating the surface retention of water. The standardized precipitation-evapotranspiration index (SPEI, Vicente-Serrano et al., 2010) has been developed to take into account the effect of evapotranspiration on surface water fluxes. Evapotranspiration is typically not calculated by climate models, so SPEI often takes other inputs to estimate it, e.g. with the Thornthwaite method based on temperature (Thornthwaite, 1948), the Hargreaves method using the monthly mean of daily minimum and maximum near-surface temperature (tasmin and tasmax) (Hargreaves, 1994), or the Penman-Monteith method using minimum and maximum temperature together with 2m wind speed (Allen et al., 1994), which is estimated from the surface wind (at 10m). However, it has been shown that the method used to derive the potential evapotranspiration has little impact on the drought statistics (Burke et al., 2006). In contrast to this finding, (Shaw and Riha, 2011) conclude that especially for future scenarios with rising temperatures potential evapotranspiration based on estimates considering temperature only can lead to an overestimation of SPEI.

In order to assess the performance of drought characteristics in climate models, three diagnostics have been implemented into the ESMValTool (v2.0): consecutive dry days, SPI and SPEI. The consecutive dry days diagnostic (*recipe_consecdrydays.yml*)



has been implemented consistently with the CDO method ‘eca_cdd’ (Climate Data Operators, Schulzweida, 2018), and the SPI and SPEI diagnostics (*recipe_spei.yml*) are based on the R-package SPEI (<https://cran.r-project.org/web/packages/SPEI/SPEI.pdf>; Vicente-Serrano et al., 2010). The recipe *recipe_spei.yml* computes the SPI and SPEI quantities for each model and summarizes the statistics of both indices as global averages in categories from “extremely dry” to “extremely wet”, see Fig. 3 and Fig. 4. By including an estimate for evapotranspiration, the model biases are reduced, particularly for the too frequent “moderately wet” category. For SPI (Fig. 3), the bias plot shows a clear underestimation of dry and wet conditions, which are mainly compensated by too frequent moderately and extremely wet conditions. For the neutral condition category, the results differ depending on the models with a tendency towards too frequent occurrence in most models. For SPEI (Fig. 4) the bias plot indicates too frequent neutral conditions, at the expense of mainly dry and wet conditions. Moderate and extreme wet conditions are overestimated in practically all models, whereas moderately and extreme dry conditions show the opposite behaviour.

Using the SPI calculation described above, a recipe analysing drought events (*recipe_martin18.yml*) has been developed. Following Martin (2018), a drought event is defined as any consecutive number of months with “extremely dry” conditions ($SPI < -2$). The characteristics of these events from historical and future scenario model runs (see Fig. 5) as well as from observational data are then compared. The characteristics investigated are frequency, length, average SPI, and the severity index following Peters (2014), which is a measure combining the length and the SPI value of a drought. Figure 5 shows an increase in the number of drought events, the severity index and to a smaller amount the duration of drought events in the RCP8.5 scenario compared to the historical model runs, especially in the subtropical areas. The results support the finding, that regions with already dry conditions are much more likely to show a higher number of drought events for the RCP 8.5 scenario, known as the “dry gets drier and the wet gets wetter” (DDWW) paradigm (Greve et al. 2014).

3.2 Extreme events

Changes in climate extremes are of utmost concern for society as the consequences of climate change will be strongly manifested in the severe impacts of extreme events, such as heatwaves and extreme precipitation, on human and natural systems. Some confidence in future projections of extreme events can be gained by evaluating the models’ performance in simulating historical events against observational data and reanalysis datasets. The 27 core climate extremes indices defined by the ETCCDI (Zhang et al., 2011) are able to capture different characteristics of temperature and precipitation extremes and are suitable for monitoring observed climate extremes, model evaluation and analysis of changes in climate extremes in future climate projections (e.g. Sillmann et al., 2013a; Sillmann et al., 2013b; Donat et al., 2013). To calculate these indices, daily values of total precipitation (pr), daily mean near-surface air temperature (tas), daily minimum near-surface air temperature (tasmin) and maximum minimum near-surface air temperature (tasmax) are required.

The *recipe_extreme_events.yml* calculates climate extremes indices and produces diagnostic figures for comparing model and observational extremes indices as presented in IPCC AR5 chapter 9 (Flato et al., 2013) and Sillmann et al. (2013a).



The index computation is performed according to Zhang et al. (2005b). The indices are calculated from CMIP models and gridded observational/reanalysis data. Calculating the indices is computationally demanding and can take hours up to days, depending on the number of models/observations, length of the time periods analysed and spatial resolution of the datasets. If possible, it is recommended to run this processing step on a parallel computing system, taking advantage of the ESMValTool task-based parallelization feature (Righi et al., 2020).

There are two types of diagnostic plots that can be produced: (i) time series providing a temporal comparison between the mean and spread (interquartile range) of the model ensemble and the individual observations for a single index (see Fig. 6); (ii) performance metrics or “portrait diagrams”, which compare multiple models with up to 4 different observations for multiple indices. The root-mean-square error (RMSE) between each model and each observational/reanalysis dataset is used as a measure for model performance (see Fig. 7). Together, the two plot types reproduce the analysis shown in figure 9.37 of IPCC AR5 (Flato et al., 2013) for a given reanalysis and model dataset. In Fig. 6, the agreement in trends between the CMIP models and reanalyses can be captured very well, due to the construction of the percentile-threshold based indices. Deviations from the nominal level of 10% outside the base period are mainly due to differences in the estimated trends in tasmin and tasmax of the individual models as compared to the respective reanalysis data set. In Sillmann et al. (2014) an alternative approach is described to evaluate percentile-threshold based indices accounting for potential model biases in the mean.

Figure 7 shows that the magnitude of the multi-model median error on a global scale (i.e. $RMSE_{std}$) is generally larger for precipitation indices than for the absolute and percentile-threshold indices based on temperature with the exception of csdi and wsdi. For the temperature-based percentile-threshold indices (i.e., tx90p, tx10p, tn90p, and tn10p), the models generally perform well (except IPSL-CM5A-LR) due to their construction. This results in good agreement for the ensemble mean and medians compared to reanalysis data, whereas the root-mean-square error is too large as it is dominated by the outlier model (IPSL-CM5A-LR).

Indices of climate extremes are a natural extension of those on the hydrological cycle discussed in Sect. 3.1 and an effort was made to make them available within the same analysis tool. As mentioned before, the ETCCDI computed by *recipe_extreme_events.yml* can be further processed by the recipe *recipe_hyint_extreme_events.yml*. Analogous to the *recipe_hyint.yml* (see also Sec. 3.1.1), it computes maps and box-averaged time series for pre-selected continental or user defined regions, computing trends and performing significance testing over the complete set of 6+27 indices. Depending on the specific objective, the user can select the needed subset of indices. Significance testing is performed with a Student’s t-test on non-null coefficients hypothesis and trend coefficients are stored together with their statistics. The recipe produces a variety of plot types for the indices, including maps and time series with their spread, trends, and summary plots of trend coefficients.



3.3 Impacts of climate change

3.3.1 Heat and cold wave duration

Heat waves are expected to become one of the greatest threats to human health in the 21st century due to projected increases in both frequency and severity (IPCC, 2013; Ouzeau et al., 2016), while the duration, intensity and frequency of cold waves are expected to decrease. It is not clear yet, however, what the impact of changes in heat and cold waves on related mortality will be, since mortality due to heat waves and cold waves inferred from historical simulations is typically overestimated. This is partly due to challenges in the correct simulation of extremes (Wang et al., 2016). In the case of heat waves in particular, models have been shown to contain biases in the 90th and 10th percentiles over the historical period (Pereira et al., 2017). However, by using a bias adjustment method based on percentiles, climate models are able to produce output which is consistent with events observed during the historical period (Ouzeau et al., 2016).

The diagnostics of the *recipe_heatwaves_coldwaves.yml* uses the daily maximum or minimum temperatures to estimate the relative change in heat as well as cold wave characteristics in future climates compared to a reference period. The user selects the model, emissions scenario, the region of interest and the reference as well as the projection periods and the percentile which will be used to compute the threshold for exceedance or non-exceedance from the reference period (a separate threshold is computed for each day of the selected season and grid point using the quantile bootstrapping method described in Zhang et al. (2005b)). Further options, which can be selected include whether to compute the frequency of exceedances or non-exceedances of extreme high or extreme low temperature events, respectively. Additionally, the minimum duration of an event to be classified as a heat/cold wave and the season of interest can be set. The diagnostic calculates the number of consecutive days over which temperature exceeds or does not exceed the given threshold in future climate projections. The result is presented as annual time series of the total number of heat or cold wave days for the selected season at each grid point and the average number of these days for the selected season in the future climate projections is calculated, see Fig. 8.

3.3.2 Combined Climate Extreme Index

High mortality rates, increases in hospital admissions as well as major economic losses are often associated with extreme events (Meehl et al., 2000; Zhang et al., 2011; Fouillet et al., 2006; Whitman et al., 1997). This emphasizes the need for monitoring and forecasting extreme events, in particular since some studies suggest that extremes are increasing in both frequency and severity with increasing anthropogenic greenhouse gases (Alexander et al., 2006; Donat et al., 2013).

The recipe *recipe_extreme_index.yml* allows a user to compute the Combined Climate Extreme Index, which is defined as a combination of different extreme values linked to precipitation, surface temperature and surface wind speed. This index is similar to the Climate Extremes Index (CEI; Karl et al., 1996), the modified CEI (mCEI; Gleason et al., 2008) or the Actuaries Climate Index (ACI; American Academy of Actuaries, 2018). In *recipe_extreme_index.yml*, the user defines the area, the reference period, the period of interest and the weights assigned for each individual component of the index. The weights allow



the user to put the emphasis on the extremes that are more relevant to them and/or completely exclude non relevant ones. Temperature and precipitation extremes are defined in a similar fashion as in Donat et al. (2013) and are part of the larger set of extreme indices compiled by the ETCCDI (Zhang et al., 2011). The different components of the multi-metric index are

- *weight_t90p*: the number of days when the maximum temperature exceeds the 90th percentile,
- *weight_t10p*: the number of days when the minimum temperature falls below the 10th percentile,
- *weight_Wx*: the number of days when wind power (third power of wind speed) exceeds the 90th percentile,
- *weight_cdd*: the maximum length of a dry spell, defined as the maximum number of consecutive days when the daily precipitation is below 1 mm, and
- *weight_rx5day*: the maximum precipitation accumulated during 5 consecutive days.

The thresholds are computed for each day in a season using a five-day running window as described in (Zhang et al., 2005a). For the calculation of the index a user-defined reference period is used for normalization and computation of the threshold corresponding to the selected metric. This recipe creates a plot containing the time average of the components listed above for the period of interest (Fig. 9a to 9e). The recipe also computes the area-weighted average of those components and combines them into a single index using the weights and the running mean (*running_mean* parameter) defined by the user. The output of the recipe consists of a netCDF file of the area-weighted and multi-model multi-metric index and a plot of the time series of that index over the selected period.

3.3.3 Daily temperature range variation

The daily temperature range (DTR) corresponds to the difference between the minimum and maximum temperature within a period of 24 hours at a given location. The usefulness of the global-average DTR has been proved using both observations and climate model simulations (Braganza et al., 2004). Changes in the mean and variability of the DTR have been shown to have a wide range of impacts on society, for example on the transmission of diseases (Lambrechts et al., 2011; Paaijmans et al., 2010) and energy consumption (Déandreis et al., 2014).

In the energy sector, a vulnerability indicator based on the DTR has been defined to identify locations which may experience increased diurnal temperature variations in the future. Increased diurnal temperature variations put additional stress on the operational management of urban heating systems. The measure is defined as the DTR exceeding 5°C at a given location and for a given day of the year (Déandreis et al., 2014). Projections of this measure are currently subject to large uncertainties as both projections of daily maximum and minimum near-surface temperature (*tasmax* and *tasmin*) in future climate projections are highly uncertain.

The recipe *recipe_diurnal_temperature_index.yml* computes the mean DTR for a given reference period using historical simulations and then the number of days on which the DTR in future climate projections exceeds that of the reference period by 5°C or more. The user can define both the reference and projection periods, and the region to be analysed. The output



produced by this recipe consists of a four-panel plot showing the maps of the projected mean DTR indicator for each season
 315 (see Fig. 10) and a netCDF file containing the corresponding data.

3.3.4 Capacity factor

The energy sector is the single biggest contributor to greenhouse gas (GHG) emissions (IPCC, 2014a). Therefore, many
 countries have adopted mitigation strategies to increase the fraction of energy generated from renewable sources in the
 forthcoming years. However, renewable energy sources like wind power and solar power rely heavily on atmospheric
 320 conditions to produce energy and are therefore exposed to risks from climate variability and long-term change in case they
 lead to detrimental atmospheric conditions. The relationship between wind speed and energy production by wind turbines is
 highly non-linear because turbines are designed to be efficient for a narrow band of wind speed conditions. Therefore, changes
 in the wind speed distribution can impact electricity generation and thus the revenues and economic viability of wind farms.
 The capacity factor is a normalized indicator of the suitability of wind speed conditions to produce electricity, irrespective of
 325 the size and number of installed turbines. The factor is provided for wind turbines designed for low, medium and high wind
 speed conditions in grouped in three different classes (IEC, 2005)..

The recipe *recipe_capacity_factor.yml* computes the wind capacity factor for these three wind turbine classes (see Fig. 11),
 taking as input the daily instantaneous surface wind speed, extrapolating to the wind speed at 100 m height as described in
 (Lledo et al., 2019). The user can select the region, period and season of interest. The result of the recipe is the capacity factor
 330 for each of the three turbine classes saved as netCDF file.

The output of solar photovoltaic (PV) systems depends on the time of the day, season, and weather conditions. The PV capacity
 factor is a measure of which fraction of the maximum possible energy is produced per grid cell. The solar power generation of
 a PV system mainly depends on the amount of incoming surface solar radiation but is also influenced by other atmospheric
 variables that affect the efficiency of PV cells, which decreases as their temperature increases. The
 335 *recipe_PVcapacity_factor.yml* computes the PV capacity factor using the daily incoming surface solar radiation and the surface
 temperature with a method described in Bett and Thornton (2016). The user can select temporal range, season, and region of
 interest.

3.4 Applications for regional scales

3.4.1 Evaluation of global climate models for selected regions

340 Climate or Earth system models with a fully coupled ocean are important tools to project the future evolution of the climate
 system in response to anthropogenic forcings, such as the increase in GHG concentrations. Despite their coarse horizontal
 resolutions (typically in the order of a hundred kilometres or less) these models can provide climate information at the regional



scale to allow for assessing the impacts of climate change. The ability of these models to simulate regional climate is an important aspect of model evaluation.

345 The recipe *recipe_flato13ipcc.yml* includes a subset of diagnostics and figures from the model evaluation chapter of the IPCC AR5 (Chapter 9, Flato et al., 2013), which compare surface parameters (such as temperature and precipitation) from models and observations at regional scales.

Mean seasonal cycle of precipitation and temperature is calculated over land areas within selected regions for individual models, the multi-model mean and observation/reanalysis data (see Fig. 12). Regional biases, including 5th, 25th, 50th, 75th
 350 and 95th percentiles of the biases, in seasonal and annual mean temperature and precipitation are evaluated for several land, polar and oceanic regions (see Figs. 13 and 14). Diagnostics allow the comparison of the multi-model mean for different projects (i.e. CMIP3, CMIP5) including information on the amplitude of the root-mean-square error. The regions used in this recipe can be irregular polygons and are defined following the IPCC Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (SREX) land regions (Seneviratne et al., 2012).

355 Systematic biases in modelled projections (Boberg and Christensen, 2012) can be investigated by ranking models against observed monthly mean temperature (see Fig. 15).

3.4.2 Stochastic Downscaling

The stochastic downscaling recipe is an example of how the ESMValTool (including its pre-processing functionalities) can be used to create a post-processing chain for further downscaling applications, but strictly speaking not a diagnostic.

360 The application of climate model projections and forecasts to impact studies at small scales, such as hydrological modelling or ecological modelling, requires to bridge the large gap between the spatial resolution of current global and regional climate models and the scales required for a correct representation of the spatial and temporal structure of precipitation at fine-scales and of the probability of extreme precipitation events. In absence of a dynamical, physically based representation, a possible approach is the use of stochastic rainfall downscaling techniques. In particular, the Rainfall Filtered AutoRegressive Model
 365 (RainFARM; Rebora et al., 2006; D'Onofrio et al., 2014; Terzago et al., 2018) method is a weather generator which has only one free parameter (which can be derived from the large scales) and which requires no further calibration. RainFARM can create ensembles of high-resolution precipitation fields from coarse scale climate model data. This method also allows quantification of uncertainties and a realistic representation of subgrid-scale variability of precipitation and of precipitation extremes, which is a crucial prerequisite for impact studies in the water sector.

370 The recipe *recipe_rainfarm.yml* allows running RainFARM within the ESMValTool. Downscaled output can be produced directly from the climate model results read by the ESMValTool and exploiting its input checking, validation and pre-processing features. The recipe produces ensembles of downscaled fields (see Fig. 16) over selected regions in netCDF format, which can then be used by users for further analysis. Notice how the downscaled fields introduce fine scale precipitation



structures, while still maintaining on average the original coarse-resolution precipitation. Different stochastic realizations are shown to demonstrate how an ensemble of realizations can be used to reproduce unresolved sub grid variability.

3.5 Multi-model ensemble member sub-selection

Large multi-model ensembles are a way to assess model and scenario uncertainties in future climate projections and other model experiments. However, considering constraints in the availability of computer time and human resources, not all available ensemble members can be included in most detailed climate impact studies associated to a given future scenario.

Therefore, despite the importance of using an ensemble that is representative for the region and process of interest covering their full uncertainty range, one or few ensemble members are often rather subjectively selected depending on, for example, their availability and simplicity to access the datasets. Using more specific information about the needs of the impact study as guidance for the selection of simulations, the resulting subset can be better suited for the purpose of climate change impact research. Here, we present an efficient and flexible tool that makes better use of the ensemble by reducing its size while maintaining important ensemble characteristics.

To find an optimal subset of significantly different model projections for a given emission scenario, a clustering algorithm is applied to the multi-model ensemble for data reduction. This technique is already used to characterize the most likely scenarios in an ensemble of weather forecasts (Ferranti and Corti, 2011; Straus et al., 2017). Similar methodologies also based on cluster analysis have been explored to select a subset from an ensemble of climate simulations (Wilcke and Barring, 2016). This approach, applied at a regional level, can also be used to identify the subset of climate model ensemble members that best represent the full range of results for further downscaling applications.

The choice of the ensemble members is made flexible in order to meet the requirements of specific (regional) climate products and can be defined according to region and user needs. The decision of which variables are considered depends on the type and goals of the climate change impact assessment. For example, a study on future hydrological floods would require in particular changes of precipitation extreme quantiles, a study on the impact of climate change on the exploitation of ski slopes would require information about changes in winter temperatures and precipitation.

EnsClus (recipe *recipe_ensclus.yml*) is a cluster analysis tool in written in Python for ensembles of climate model simulations. The tool is based on the k-means algorithm with the aim to group ensemble members by similar characteristics and to select the most representative member for each cluster. The user chooses which characteristic is used to group the ensemble members by the clustering: maximum, a given percentile (75% in the examples below), mean, standard deviation or trend over the period. For each ensemble member this value is computed at each grid point. This results in N latitude-longitude maps, where N is the number of ensemble members. The anomalies are computed by subtracting the ensemble mean of these maps from each of the individual maps. The anomalies are therefore not computed with respect to time but to the ensemble members. An Empirical Orthogonal Function (EOF) analysis is performed on these anomaly maps. For the EOF analysis, the user can set either how many Principal Components (PCs) should be calculated or the minimum percentage of the explained variance



which should be covered. After reducing dimensionality via EOF analysis, the k-means algorithm is applied using the selected PCs (the number k of clusters needs to be defined prior to the analysis). The output of the recipe is a classification by clusters, i.e. which ensemble member belongs to which cluster and the most representative ensemble member for each cluster, defined by the member being closest to the cluster centroid. Additionally, output of the recipe includes the statistics of clustering: in the PC space, the minimum and the maximum distance between a member in a cluster and the cluster centroid (i.e. the closest and the farthest member), and the intra-cluster standard deviation for each cluster (i.e. compactness of the cluster). An example is shown in Fig. 17. The figure shows a clustering based on the 75th percentile of historical summer (JJA) precipitation rate for 32 CMIP5 models over 1900-2005. Based on the principal components explaining 80% of the variance three clusters are computed. The green cluster is the most populated with 16 ensemble members. It is mostly characterized by a positive anomaly over central-north Europe. The red cluster contains 12 ensemble members. It exhibits a negative anomaly centred over southern Europe and in few cases (e.g. No.12 and No.23) extending north. The third cluster (blue) includes only 4 models. It is showing a north-south dipolar precipitation anomaly, with a wetter than average Mediterranean counteracting dryer North-Europe. Ensemble members No.9, No.26 and No.19 are the “specimen” of each cluster, i.e. the model simulations that best represent the main features of that cluster. These three ensemble members can eventually be used as representative of the whole possible outcomes of the multi-model ensemble distribution associated to the 32 CMIP5 historical integrations for the summer precipitation rate 75th percentile over Europe. This reduces the outcomes from 32 to 3 ensemble members. The number of ensemble members of each cluster might provide a measure of the probability of occurrence of each cluster. However, the final results are sensitive to models’ bias and to the metric used, as in any selection exercise.

4. Summary

This paper summarizes the recipes available within the ESMValTool v2.0 for the analysis of extreme events, droughts, model impact assessment, sub-selection of multi-model ensemble members e.g. for downscaling applications, as well as model evaluation on regional scales. It complements the series of papers that have been published on ESMValTool v2.0 by Righi et al. (2020) describing the technical aspects of ESMValTool v2.0, Eyring et al. (2020) presenting the new large-scale diagnostics that have been included in v2.0 since the first release in 2016 (Eyring et al., 2016), and Lauer et al. (2020) covering emergent constraints and diagnostics for the analysis of future projections from ESMs in CMIP.

For droughts, recipes calculating the consecutive number of dry days, the SPI, and the SPEI have been newly included in ESMValTool v2.0 as well as a recipe to analyse the frequency, length, and severity of drought events based on the SPI. For further analysis of extreme events, climate extreme indices of the Expert Team on Climate Change Detection and Indices (ETCCDI) based on Zhang et al. (2011) have been included. These indices are calculated based on daily total precipitation, and the mean, minimum and maximum of the near-surface air temperature. The indices can then be plotted, used as a measure of model performance, and further processed to calculate index trends and their significance.



For model impact assessments, recipes to analyse heat and cold wave duration, diurnal temperature variations, as well as different extreme indices are included in ESMValTool v2.0. Additional recipes compute capacity factors to analyse the impact of climate change on the wind and solar energy production.

440 For the analysis of ensembles of climate models, ESMValTool v2.0 provides a cluster analysis based on a k-means algorithm where the ensemble members are divided into clusters and can be plotted along with the properties of the clusters and the most representative member of each cluster.

ESMValTool v2.0 also includes diagnostics for model evaluation on regional scales. Surface parameters such as temperature and precipitation can be evaluated for regions defined by polygons following the SPEX definitions of land regions.

445 Additionally, the ESMValTool output can be used to be processed further by tools for stochastic downscaling, like RainFARM which is also implemented in v2.0.

Although the recipes here are presented using CMIP5 data, ESMValTool v2.0 can be run to perform the same analysis for CMIP6 data. As an open source project, the capabilities of the ESMValTool continue to grow with contributions from the scientific community being highly welcome. Users can analyse data using a wealth of existing recipes or join the ESMValTool

450 development team and add new recipes and diagnostics.

5. Code availability and data availability

Code and data availability. ESMValTool v2.0 is released under the Apache License, VERSION 2.0. The latest release of ESMValTool v2.0 is publicly available on Zenodo at <http://doi.org/10.5281/zenodo.3970975> (Andela et al., 2020a). The source code of the ESMValCore package, which is installed as a dependency of the ESMValTool v2.0, is also publicly available on
455 Zenodo at <http://doi.org/10.5281/zenodo.3952695> (Andela et al., 2020b). ESMValTool and ESMValCore are developed on the GitHub repositories available at <https://github.com/ESMValGroup> (last access: 24 July 2020). CMIP5 data are available freely and publicly from the Earth System Grid Federation. Observations used in the evaluation are detailed in the various sections of the paper and listed in Table 1. They are not distributed with ESMValTool, which is restricted to the code as open-source software.

460 Author contribution

KW led the writing of the paper and with the help of LB, BKG AL, MR, and MS coordinated the implementation of the diagnostics for this paper in ESMValTool v2.0. VE coordinated the ESMValTool v2.0 release. All other authors contributed to individual diagnostics for this release. All authors contributed to the text.



465 Acknowledgements

The diagnostic development of ESMValTool v2.0 for this paper was supported by different projects with different scientific focus, in particular by: (1) Copernicus Climate Change Service (C3S) "Metrics and Access to Global Indices for Climate Projections (C3S-MAGIC)" project C3S_34a Lot 2, (2) Horizon 2020 European Union's Framework Programme for Research and Innovation under Grant Agreement No. 641816, project CRESCENDO (Coordinated Research in Earth Systems and
 470 Climate: Experiments, kNowledge, Dissemination and Outreach), (3) Helmholtz Society project "Advanced Earth System Model Evaluation for CMIP (EVal4CMIP)", and (4) Federal Ministry of Education and Research (BMBF) CMIP6-DICAD project. In addition, we received technical support on the ESMValTool v2.0 development from the European Union's Horizon 2020 Framework Programme for Research and Innovation "Infrastructure for the European Network for Earth System Modelling (IS-ENES3)" project under Grant Agreement No 824084. We acknowledge the World Climate Research Program's
 475 (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. This work used JASMIN, the UK collaborative data analysis facility. The computational resources of the Deutsches Klimarechenzentrum (DKRZ, Germany) were also essential for developing and testing this new version and are kindly acknowledged.

480 References

- Alexander, L. V., Zhang, X., Peterson, T. C., Caesar, J., Gleason, B., Tank, A. M. G. K., Haylock, M., Collins, D., Trewin, B., Rahimzadeh, F., Tagipour, A., Kumar, K. R., Revadekar, J., Griffiths, G., Vincent, L., Stephenson, D. B., Burn, J., Aguilar, E., Brunet, M., Taylor, M., New, M., Zhai, P., Rusticucci, M., and Vazquez-Aguirre, J. L.: Global observed changes in daily climate extremes of temperature and precipitation, *J Geophys Res-Atmos*, 111, Artn D05109
 485 10.1029/2005jd006290, 2006.
- Alexander, L. V.: Global observed long-term changes in temperature and precipitation extremes: A review of progress and limitations in IPCC assessments and beyond, *Weather Clim Extreme*, 11, 4-16, 10.1016/j.wace.2015.10.007, 2016.
- Allen, R. G., Smith, M., Pereira, L. S., and Perrier, A.: An update for the calculation of reference evapotranspiration., *ICID Bulletin of the International Commission on Irrigation and Drainage*, 35-92, 1994.
- 490 Actuaries Climate Index: <http://actuariesclimateindex.org>, access: 2018-10-06, 2018.
- Andela, B., Broetz, B., de Mora, L., Drost, N., Eyring, V., Koldunov, N., Lauer, A., Mueller, B., Predoi, V., Righi, M., Schlund, M., Vegas-Regidor, J., Zimmermann, K., Adeniyi, K., Arnone, E., Bellprat, O., Berg, P., Bock, L., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Corti, S., Crezee, B., Davin, E. L., Davini, P., Deser, C., Diblen, F., Docquier, D., Dreyer, L., Ehbrecht, C., Earnshaw, P., Gier, B., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., von Hardenberg, J., Hassler, B., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Lledó, L., Lejeune, Q., Lembo, V., Little, B., Loosveldt-Tomas, S., Lorenz, R., Lovato, T., Lucarini, V., Massonnet, F., Mohr, C. W., Amarjiit, P., Pérez-Zanón, N., Phillips, A., Russell, J., Sandstad, M., Sellar, A., Senfleben, D., Serva, F., Sillmann, J., Stacke, T., Swaminathan, R., Torralba, V., and Weigel, K.: ESMValTool (Version v2.0.0b4), Zendo, <http://doi.org/10.5281/zenodo.3759523>, 2020a.



- Andela, B., Broetz, B., de Mora, L., Drost, N., Eyring, V., Koldunov, N., Lauer, A., Predoi, V., Righi, M., Schlund, M., Vegas-Regidor, J., Zimmermann, K., Bock, L., Diblen, F., Dreyer, L., Earnshaw, P., Hassler, B., Little, B., and Loosveldt-Tomas, S.: ESMValCore (Version v2.0.0), Zenodo, <http://doi.org/10.5281/zenodo.3952695>, 2020b.
- Bett, P. E., and Thornton, H. E.: The climatological relationships between wind and solar energy supply in Britain, *Renew Energ*, 87, 96–110, 10.1016/j.renene.2015.10.006, 2016.
- Boberg, F., and Christensen, J. H.: Overestimation of Mediterranean summer temperature projections due to model deficiencies, *Nature Climate Change*, 2, 433–436, 10.1038/Nclimate1454, 2012.
- 505 Braganza, K., Karoly, D. J., and Arblaster, J. M.: Diurnal temperature range as an index of global climate change during the twentieth century, *Geophysical Research Letters*, 31, n/a–n/a, 10.1029/2004gl019998, 2004.
- Burke, E. J., Brown, S. J., and Christidis, N.: Modeling the recent evolution of global drought and projections for the twenty-first century with the hadley centre climate model, *J Hydrometeorol*, 7, 1113–1125, Doi 10.1175/Jhm544.1, 2006.
- Cheng, L. Y., Hoerling, M., AghaKouchak, A., Livneh, B., Quan, X. W., and Eischeid, J.: How Has Human-Induced Climate Change
 510 Affected California Drought Risk?, *J Climate*, 29, 111–120, 10.1175/Jcli-D-15-0260.1, 2016.
- D'Onofrio, D., Palazzi, E., von Hardenberg, J., Provenzale, A., and Calmanti, S.: Stochastic Rainfall Downscaling of Climate Models, *J Hydrometeorol*, 15, 830–843, 10.1175/Jhm-D-13-096.1, 2014.
- Impact du changement climatique sur la gestion des réseaux de chaleur. : <https://docplayer.fr/9496504-Impact-du-changement-climatique-sur-la-gestion-des-reseaux-de-chaleur.html>, access: 11.02.2020, 2014.
- 515 Deser, C., Phillips, A., Bourdette, V., and Teng, H. Y.: Uncertainty in climate change projections: the role of internal variability, *Clim Dynam*, 38, 527–546, 10.1007/s00382-010-0977-x, 2012.
- Donat, M. G., Alexander, L. V., Yang, H., Durre, I., Vose, R., Dunn, R. J. H., Willett, K. M., Aguilar, E., Brunet, M., Caesar, J., Hewitson, B., Jack, C., Tank, A. M. G. K., Kruger, A. C., Marengo, J., Peterson, T. C., Renom, M., Rojas, C. O., Rusticucci, M., Salinger, J., Elrayah, A. S., Sekele, S. S., Srivastava, A. K., Trewin, B., Villarreal, C., Vincent, L. A., Zhai, P., Zhang, X., and Kitching, S.: Updated analyses of
 520 temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset, *J Geophys Res-Atmos*, 118, 2098–2118, 10.1002/jgrd.50150, 2013.
- Donat, M. G., Angelil, O., and Ukkola, A. M.: Intensification of precipitation extremes in the world's humid and water-limited regions, *Environmental Research Letters*, 14, ARTN 065003 10.1088/1748-9326/ab1c8e, 2019.
- 525 Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C., Anav, A., Andrews, O., Cionni, I., Davin, E. L., Deser, C., Ehbrecht, C., Friedlingstein, P., Gleckler, P., Gottschaldt, K. D., Hagemann, S., Juckes, M., Kindermann, S., Krasting, J., Kunert, D., Levine, R., Loew, A., Makela, J., Martin, G., Mason, E., Phillips, A. S., Read, S., Rio, C., Roehrig, R., Senftleben, D., Sterl, A., van Ulft, L. H., Walton, J., Wang, S. Y., and Williams, K. D.: ESMValTool (v1.0) - a community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP, *Geosci Model Dev*, 9, 1747–1802, 10.5194/gmd-9-1747-2016, 2016.
- 530 Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Brötz, B., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Crezee, B., Davin, E., Davini, P., Debeire, K., de Mora, L., Deser, C., Docquier, D., Earnshaw, P., Ehbrecht, C., Gier, B. K., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hardiman, S., Hassler, B., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Koldunov, N. V., Lejeune, Q., Lembo, V., Lovato, T., Lucarini, V., Massonnet, F., Müller, B., Pandde, A., Pérez-Zanón, N., Phillips, A., Predoi, V., Russell, J., Sellar, A., Serva, F., Stacke, T., Swaminathan, R., Torralba, V., Vegas-Regidor, J., von Hardenberg, J., Weigel, K.,
 535 and Zimmermann, K.: ESMValTool v2.0 – Extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP, *Geosci. Model Dev.*, in press, <https://doi.org/10.5194/gmd-2019-291>, 2020.
- New clustering products: <https://www.ecmwf.int/en/elibrary/14596-newsletter-no-127-spring-2011>, access: 2019-10-07, 2011.



- 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, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 741–866, 2013.
- 540 Fouillet, A., Rey, G., Laurent, F., Pavillon, G., Bellec, S., Guihenneuc-Jouyaux, C., Clavel, J., Jouglu, E., and Hemon, D.: Excess mortality related to the August 2003 heat wave in France, *Int Arch Occ Env Hea*, 80, 16–24, 10.1007/s00420-006-0089-4, 2006.
- 545 Giorgi, F., Im, E. S., Coppola, E., Diffenbaugh, N. S., Gao, X. J., Mariotti, L., and Shi, Y.: Higher Hydroclimatic Intensity with Global Warming, *J Climate*, 24, 5309–5324, 10.1175/2011jcli3979.1, 2011.
- Giorgi, F., Coppola, E., and Raffaele, F.: A consistent picture of the hydroclimatic response to global warming from multiple indices: Models and observations, *J Geophys Res-Atmos*, 119, 11695–11708, 10.1002/2014jd022238, 2014.
- Giorgi, F., Raffaele, F., and Coppola, E.: The response of precipitation characteristics to global warming from climate projections, *Earth Syst Dynam*, 10, 73–89, 10.5194/esd-10-73-2019, 2019.
- 550 Gleason, K. L., Lawrimore, J. H., Levinson, D. H., Karl, T. R., and Karoly, D. J.: A revised US Climate Extremes Index, *J Climate*, 21, 2124–2137, 10.1175/2007jcli1883.1, 2008.
- Greve, P., Orlowsky, B., Mueller, B., Sheffield, J., Reichstein, M., and Seneviratne, S. I.: Global assessment of trends in wetting and drying over land, *Nature Geoscience*, 7, 716–721, 10.1038/ngeo2247, 2014.
- 555 Hargreaves, G. H.: Defining and Using Reference Evapotranspiration, *Journal of Irrigation and Drainage Engineering*, 120, 1132–1139, doi:10.1061/(ASCE)0733-9437(1994)120:6(1132), 1994.
- IPCC: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change, Cambridge, United Kingdom and New York, NY, USA, 582 pp, 2012.
- 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, 1535 pp., 2013.
- 560 IPCC: Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge, United Kingdom and New York, NY, USA, 2014a.
- IPCC: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, 1132 pp. , 2014b.
- 565 Karl, T. R., Knight, R. W., Easterling, D. R., and Quayle, R. G.: Indices of climate change for the United States, *B Am Meteorol Soc*, 77, 279–292, Doi 10.1175/1520-0477(1996)077<0279:Ioccf>2.0.Co;2, 1996.
- Lambrechts, L., Paaijmans, K. P., Fansiri, T., Carrington, L. B., Kramer, L. D., Thomas, M. B., and Scott, T. W.: Impact of daily temperature fluctuations on dengue virus transmission by *Aedes aegypti*, *Proceedings of the National Academy of Sciences*, 108, 7460–7465, 10.1073/pnas.1101377108, 2011.
- 570 Lauer, A., Eyring, V., Bellprat, O., Bock, L., Gier, B. K., Hunter, A., Lorenz, R., Pérez-Zanón, N., Righi, M., Schlund, M., Senftleben, D., Weigel, K., and Zechlau, S.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – diagnostics for emergent constraints and future projections from Earth system models in CMIP, *Geosci. Model Dev. Discuss.*, accepted, 2020, 1–47, 10.5194/gmd-2020-60, 2020.



- 575 Lledo, L., Torralba, V., Soret, A., Ramon, J., and Doblas-Reyes, F. J.: Seasonal forecasts of wind power generation, *Renew Energ*, 143, 91-100, 10.1016/j.renene.2019.04.135, 2019.
- Maraun, D., Shepherd, T. G., Widmann, M., Zappa, G., Walton, D., Gutierrez, J. M., Hagemann, S., Richter, I., Soares, P. M. M., Hall, A., and Mearns, L. O.: Towards process-informed bias correction of climate change simulations, *Nature Climate Change*, 7, 764-773, 10.1038/Nclimate3418, 2017.
- 580 Martin, E. R.: Future Projections of Global Pluvial and Drought Event Characteristics, *Geophysical Research Letters*, 45, 11913-11920, 10.1029/2018gl079807, 2018.
- Marvel, K., Cook, B. I., Bonfils, C. J. W., Durack, P. J., Smerdon, J. E., and Williams, A. P.: Twentieth-century hydroclimate changes consistent with human influence, *Nature*, 569, 59-+, 10.1038/s41586-019-1149-8, 2019.
- 585 McKee, T. B., Doesken, N. J., and Kleist, J.: The relationship of drought frequency and duration to time scales., *Proceedings of the 8th Conference on Applied Climatology*, American Meteorological Society Boston, MA, 179-183, 1993.
- Meehl, G. A., Karl, T., Easterling, D. R., Changnon, S., Pielke, R., Changnon, D., Evans, J., Groisman, P. Y., Knutson, T. R., Kunkel, K. E., Mearns, L. O., Parmesan, C., Pulwarty, R., Root, T., Sylves, R. T., Whetton, P., and Zwiers, F.: An introduction to trends in extreme weather and climate events: Observations, socioeconomic impacts, terrestrial ecological impacts, and model projections, *B Am Meteorol Soc*, 81, 413-416, Doi 10.1175/1520-0477(2000)081<0413:Aittie>2.3.Co;2, 2000.
- 590 Ouzeau, G., Soubeyroux, J. M., Schneider, M., Vautard, R., and Planton, S.: Heat waves analysis over France in present and future climate: Application of a new method on the EURO-CORDEX ensemble, *Climate Services*, 4, 1-12, 10.1016/j.cliser.2016.09.002, 2016.
- Paaijmans, K. P., Blanford, S., Bell, A. S., Blanford, J. I., Read, A. F., and Thomas, M. B.: Influence of climate on malaria transmission depends on daily temperature variation, *Proceedings of the National Academy of Sciences*, 107, 15135-15139, 10.1073/pnas.1006422107, 2010.
- 595 Palmer, W. C.: Meteorologic drought, *U.S. Weather Bureau, Res. Pap.*, 45, 58, 1965.
- Pereira, S. C., Marta-Almeida, M., Carvalho, A. C., and Rocha, A.: Heat wave and cold spell changes in Iberia for a future climate scenario, *International Journal of Climatology*, 37, 5192-5205, 10.1002/joc.5158, 2017.
- Peters, E. J.: Measuring the Severity of Dry Seasons in the Grenadines, *The West Indian Journal of Engineering*, 36, 42-50, 2014.
- 600 Pfahl, S., O'Gorman, P. A., and Fischer, E. M.: Understanding the regional pattern of projected future changes in extreme precipitation, *Nature Climate Change*, 7, 423-+, 10.1038/Nclimate3287, 2017.
- Rebora, N., Ferraris, L., von Hardenberg, J., and Provenzale, A.: RainFARM: Rainfall downscaling by a filtered autoregressive model, *J Hydrometeorol*, 7, 724-738, Doi 10.1175/Jhm517.1, 2006.
- 605 Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brötz, B., de Mora, L., Diblen, F., Dreyer, L., Drost, N., Earnshaw, P., Hassler, B., Koldunov, N., Little, B., Loosveldt Tomas, S., and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview, *Geosci. Model Dev.*, 13, 1179-1199, 10.5194/gmd-13-1179-2020, 2020.
- 610 Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., and Zhang, X.: Changes in climate extremes and their impacts on the naturalphysical environment. , in: *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on ClimateChange (IPCC)*, edited by: Field, C. B., Barros, V., Stocker, T. F., Qin, D., Dokken, D. J., Ebi, K. L., Mastrandrea, M. D., Mach, K. J., Plattner, G.-K., Allen, S. K., Tignor, M., and Midgley, P. M., Cambridge University Press, Cambridge, UK, and New York, NY, USA, 109-230, 2012.



- Shaw, S. B., and Riha, S. J.: Assessing temperature-based PET equations under a changing climate in temperate, deciduous forests, *Hydrol Process*, 25, 1466-1478, 10.1002/hyp.7913, 2011.
- 615 Shepherd, T. G.: Atmospheric circulation as a source of uncertainty in climate change projections, *Nature Geoscience*, 7, 703-708, 10.1038/Ngeo2253, 2014.
- Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., and Bronaugh, D.: Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate, *J Geophys Res-Atmos*, 118, 1716-1733, 10.1002/jgrd.50203, 2013a.
- Sillmann, J., Kharin, V. V., Zwiers, F. W., Zhang, X., and Bronaugh, D.: Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections, *J Geophys Res-Atmos*, 118, 2473-2493, 10.1002/jgrd.50188, 2013b.
- 620 Sillmann, J., Kharin, V. V., Zwiers, F. W., Zhang, X., Bronaugh, D., and Donat, M. G.: Short Communication Evaluating model-simulated variability in temperature extremes using modified percentile indices, *International Journal of Climatology*, 34, 3304-3311, 10.1002/joc.3899, 2014.
- Straus, D., Molteni, F., and Corti, S.: Atmospheric Regimes: The Link between Weather and the Large-Scale Circulation., in: *Nonlinear and Stochastic Climate Dynamics*, edited by: Franzke, C., and O'Kane, T., Cambridge University Press., Cambridge, 105-135, 2017.
- 625 Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of Cmp5 and the Experiment Design, *B Am Meteorol Soc*, 93, 485-498, 10.1175/Bams-D-11-00094.1, 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, 10.1175/Bams-D-12-00204.1, 2014.
- 630 Terzago, S., Palazzi, E., and von Hardenberg, J.: Stochastic downscaling of precipitation in complex orography: a simple method to reproduce a realistic fine-scale climatology, *Nat Hazard Earth Sys*, 18, 2825-2840, 10.5194/nhess-18-2825-2018, 2018.
- Thornthwaite, C. W.: An Approach toward a Rational Classification of Climate, *Geogr Rev*, 38, 55-94, Doi 10.2307/210739, 1948.
- Ukkola, A. M., Pitman, A. J., De Kauwe, M. G., Abramowitz, G., Herger, N., Evans, J. P., and Decker, M.: Evaluating CMIP5 Model Agreement for Multiple Drought Metrics, *J Hydrometeorol*, 19, 969-988, 10.1175/Jhm-D-17-0099.1, 2018.
- 635 Vicente-Serrano, S. M., Begueria, S., and Lopez-Moreno, J. I.: A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index, *J Climate*, 23, 1696-1718, 10.1175/2009jcli2909.1, 2010.
- Waliser, D., Gleckler, P. J., Ferraro, R., Taylor, K. E., Ames, S., Biard, J., Bosilovich, M. G., Brown, O., Chepfer, H., Cinquini, L., Durack, P. J., Eyering, V., Mathieu, P.-P., Lee, T., Pinnock, S., Potter, G. L., Rixen, M., Saunders, R., Schulz, J., Thépaut, J.-N., and Tuma, M.: Observations for Model Intercomparison Project (Obs4MIPs): status for CMIP6, *Geosci Model Dev*, 13, 2945-2958, 10.5194/gmd-13-2945-2020, 2020.
- 640 Wang, Y., Shi, L., Zanobetti, A., and Schwartz, J. D.: Estimating and projecting the effect of cold waves on mortality in 209 US cities, *Environment International*, 94, 141-149, 10.1016/j.envint.2016.05.008, 2016.
- Watts, N., Adger, W. N., Agnolucci, P., Blackstock, A., Byass, P., Cai, W. J., Chaytor, S., Colbourn, T., Collins, M., Cooper, A., Cox, P. M., Depledge, J., Drummond, P., Ekins, P., Galaz, V., Grace, D., Graham, H., Grubb, M., Haines, A., Hamilton, I., Hunter, A., Jiang, X. J., Li, M. X., Kelman, I., Liang, L., Lott, M., Lowe, R., Luo, Y., Mace, G., Maslin, M., Nilsson, M., Oreszczyn, T., Pye, S., Quinn, T., Svensdotter, M., Venevsky, S., Warner, K., Xu, B., Yang, J., Yin, Y. Y., Yu, C. Q., Zhang, Q., Gong, P., Montgomery, H., and Costello, A.: Health and climate change: policy responses to protect public health, *Lancet*, 386, 1861-1914, 10.1016/S0140-6736(15)60854-6, 2015.
- 645 Whitman, S., Good, G., Donoghue, E. R., Benbow, N., Shou, W. Y., and Mou, S. X.: Mortality in Chicago attributed to the July 1995 heat wave, *Am J Public Health*, 87, 1515-1518, Doi 10.2105/Ajph.87.9.1515, 1997.



- 650 Wilcke, R. A. I., and Barring, L.: Selecting regional climate scenarios for impact modelling studies, *Environ Modell Softw*, 78, 191-201, 10.1016/j.envsoft.2016.01.002, 2016.
- Xie, S. P., Deser, C., Vecchi, G. A., Collins, M., Delworth, T. L., Hall, A., Hawkins, E., Johnson, N. C., Cassou, C., Giannini, A., and Watanabe, M.: Towards predictive understanding of regional climate change, *Nature Climate Change*, 5, 921-930, 10.1038/nclimate2689, 2015.
- 655 Zhang, X. B., Aguilar, E., Sensoy, S., Melkonyan, H., Tagiyeva, U., Ahmed, N., Kotaladze, N., Rahimzadeh, F., Taghipour, A., Hantosh, T. H., Albert, P., Semawi, M., Ali, M. K., Al-Shabibi, M. H. S., Al-Oulan, Z., Zafari, T., Khelet, I. A., Hamoud, S., Sagir, R., Demircan, M., Eken, M., Adiguzel, M., Alexander, L., Peterson, T. C., and Wallis, T.: Trends in Middle East climate extreme indices from 1950 to 2003, *J Geophys Res-Atmos*, 110, Artid D22104 10.1029/2005jd006181, 2005a.
- 660 Zhang, X. B., Hegerl, G., Zwiers, F. W., and Kenyon, J.: Avoiding inhomogeneity in percentile-based indices of temperature extremes, *J Climate*, 18, 1641-1651, Doi 10.1175/Jcli3366.1, 2005b.
- Zhang, X. B., Alexander, L., Hegerl, G. C., Jones, P., Tank, A. K., Peterson, T. C., Trewin, B., and Zwiers, F. W.: Indices for monitoring changes in extremes based on daily temperature and precipitation data, *Wires Clim Change*, 2, 851-870, 10.1002/wcc.147, 2011.
- 665 Zscheischler, J., Westra, S., van den Hurk, B. J. J. M., Seneviratne, S. I., Ward, P. J., Pitman, A., AghaKouchak, A., Bresch, D. N., Leonard, M., Wahl, T., and Zhang, X. B.: Future climate risk from compound events, *Nature Climate Change*, 8, 469-477, 10.1038/s41558-018-0156-3, 2018.



Table 1. Overview of recipes implemented in ESMValTool v2.0 along with the section they are described, a brief description, the variables used, and the diagnostic scripts included. For further details, we refer to the GitHub repository and documentation at <https://docs.esmvaltool.org/>.

Recipe name	Section (Figures)	Description, References	Variables (Observational datasets)	Diagnostic scripts
Section 3.1: Hydrological cycle				
<i>recipe_hyint.yml</i>	3.1.1 (Fig 1)	Recipe for evaluating the intensity of the hydroclimatic cycle, calculating a set of 6 indices following Giorgi et al. (2011, 2014): simple precipitation intensity index (SDII), maximum dry spell length (DSL) and wet spell length (WSL), hydroclimatic intensity index (HY-INT), which is a measure of the overall behaviour of the hydroclimatic cycle, and precipitation area (PA), i.e. the area over which on any given day precipitation occurs.	pr	hyint/hyint.R
<i>recipe_hyint_extreme_events.yml</i>	3.1.1 (Fig 2)	Multi-diagnostic version of hyint, which allows to include ETCCDI results from the extreme_events diagnostics and performs joint analysis of indices for hydroclimatic intensity and extreme events. Giorgi et al. (2014); Giorgi et al. (2011); (Sillmann et al. 2013a)	pr tasmin tasmax	hyint/hyint.R extreme_events/extreme_events.R



<i>recipe_consecdrydays.yml</i>	3.1.2	Dry day definition (precip limit mm/day) and drought duration (days) can be set by the user. Output as netCDF files for each model possible. Computed consistently with the CDO method “eca_cdd” in Schulzweida (2018)	pr	droughtindex/diag_cdd.py
<i>recipe_spei.yml</i>	3.1.2 (Figs. 3, 4)	Global average histogram of SPI and SPEI, as absolute values and as bias. The calculations are based on pr for both indices, but for SPEI with the additional use of ta to derive evapotranspiration using the Thornthwaite method. Requires a reference dataset and calculates a global cosine of latitude weighted histogram for all valid grid points of the reference data set. Calculation of SPI and SPEI based on Vicente-Serrano et al. (2010)	pr (ERA-Interim) ta (CRUs4.01)	droughtindex/diag_spi.r droughtindex/diag_spei.r



<i>recipe_martin18grl.yml</i>	3.1.2 (Fig. 5)	Computes a monthly time series of SPI, based on <code>diag_spi.r</code> (distribution and representing time scale can be set by the user) and calculates drought events as consecutive number of months with $SPI < -2$. For each grid point the drought characteristics (frequency, average duration, and SPI as well as the severity index) based on Martin (2018) are calculated. Differences between individual models or a multi model mean and observations or future scenarios and historical model runs are calculated.	pr (CRU)	droughtindex/diag_save_spi.R droughtindex/collect_drought_func.py droughtindex/collect_drought_obs_multi.py droughtindex/collect_drought_model.py
Section 3.2: Extreme events				
<i>recipe_extreme_events.yml</i>	3.2 (Figs. 6, 7)	Calculate indices for monitoring changes in extremes (Sillmann et al., 2013a) based on daily temperature and precipitation data. Produces Glecker and time series plots as shown in the IPCC AR5 report (Flato et al., 2013)	pr tas tasmin tasmax (ERA-Interim)	extreme_events/extreme_events.R
Section 3.3: Evaluation for impact assessments				



<i>recipe_heatwaves_coldwaves.yml</i>	3.3.1 (Fig. 8)	MAGIC, time averages, differences between historical simulations and a future scenario, calculates the number of days exceeding a given quantile for a minimum number of consecutive days. Watts et al. (2015)	Tasmin Tasmax	magic_bsc/extreme_spells.r
<i>recipe_extreme_index.yml</i>	3.3.2 (Fig. 9)	MAGIC, computes time series of the number of several extreme events: heatwave, cold wave, heavy precipitation, drought, and high wind. Karl et al. (1996); Gleason et al. (2008); American Academy of Actuaries (2018)	Tasmin tasmax pr scfWind	magic_bsc/extreme_index.r
<i>recipe_diurnal_temperature_index.yml</i>	3.3.3 (Fig. 10)	MAGIC, time averages, difference between historical and future scenario, computes the dates where the DTR exceeds a threshold, Déandreis et al. (2014)	Tasmin Tasmax	magic_bsc/diurnal_temp_index.r
<i>recipe_capacity_factor.yml</i>	3.3.4 (Fig. 11)	MAGIC, calculates the wind power capacity factor, Lledo et al. (2019)	scfWind	magic_bsc/capacity_factor.r
Section 3.4: Regional model evaluation				



<i>recipe_flato13ipcc.yml</i>	3.4.1 (Figs. 12, 13, 14, 15)	Figures similar to figures of the IPCC AR5 (Flato et al., 2013). Fig. 12: Seasonal cycle over land within defined regions (like Fig. 9.38) Fig 13: Downscaling: Seasonal bias box plot within defined regions (like Fig. 9.39) Fig. 14: Downscaling: Seasonal bias box plot within defined polar and ocean regions (like Fig. 9.40) Fig. 15: Downscaling: observations versus models within defined regions (like Fig. 9.41)	tas (ERA-Interim, CRU) pr (CRU)	cfg_ipcc_ch9/cfg_Figure_9.38.ncl cfg_ipcc_ch9/cfg_Figure_9.39.ncl cfg_ipcc_ch9/cfg_Figure_9.40.ncl cfg_ipcc_ch9/cfg_Figure_9.41.ncl
<i>recipe_rainfarm.yml</i>	3.4.2 (Fig. 16)	MAGIC, Stochastic spatial downscaling of daily precipitation using the RainFARM method (Rebora et al. 2006; D'Onofrio et al. 2014). Allows calculation of climatological weights to take into account the effect of orography following Terzago et al. (2018). Produces ensembles of downscaled precipitation fields in netCDF format. No plots are produced.	pr	rainfarm/rainfarm.R

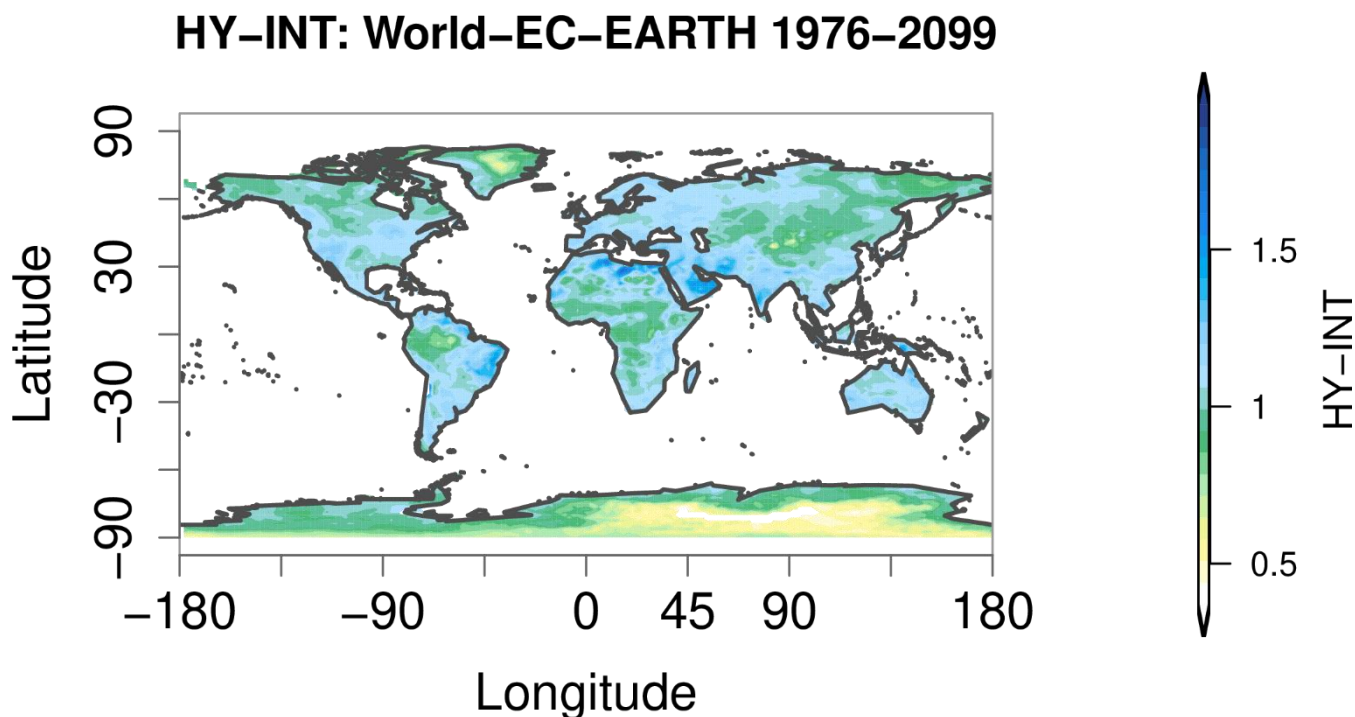
Section 3.5: Multi-model ensemble member sub-selection



<i>recipe_ensclus.yml</i>	3.5 (Fig. 17)	Cluster analysis tool for ensembles of climate model simulations. EnsClus groups ensemble members according to similar characteristics (based on the k-means algorithm) and selects the most representative member for each cluster (Straus et al. (2007))	Pr tas	ensclus/ensclus.py
---------------------------	---------------	--	-----------	--------------------



675



680 **Figure 1: Mean hydroclimatic intensity (i.e., a combination of precipitation intensity and dry spell length) for the EC-EARTH model historical simulation + RCP 8.5 projection over 1976-2099. Data were normalized to the historical 1976-2005 period. The figure is an examples of a large number of different plots which can be produced with *recipe_hyint.yml*, similar to (Giorgi et al., 2014). For details see Section 3.1.1.**

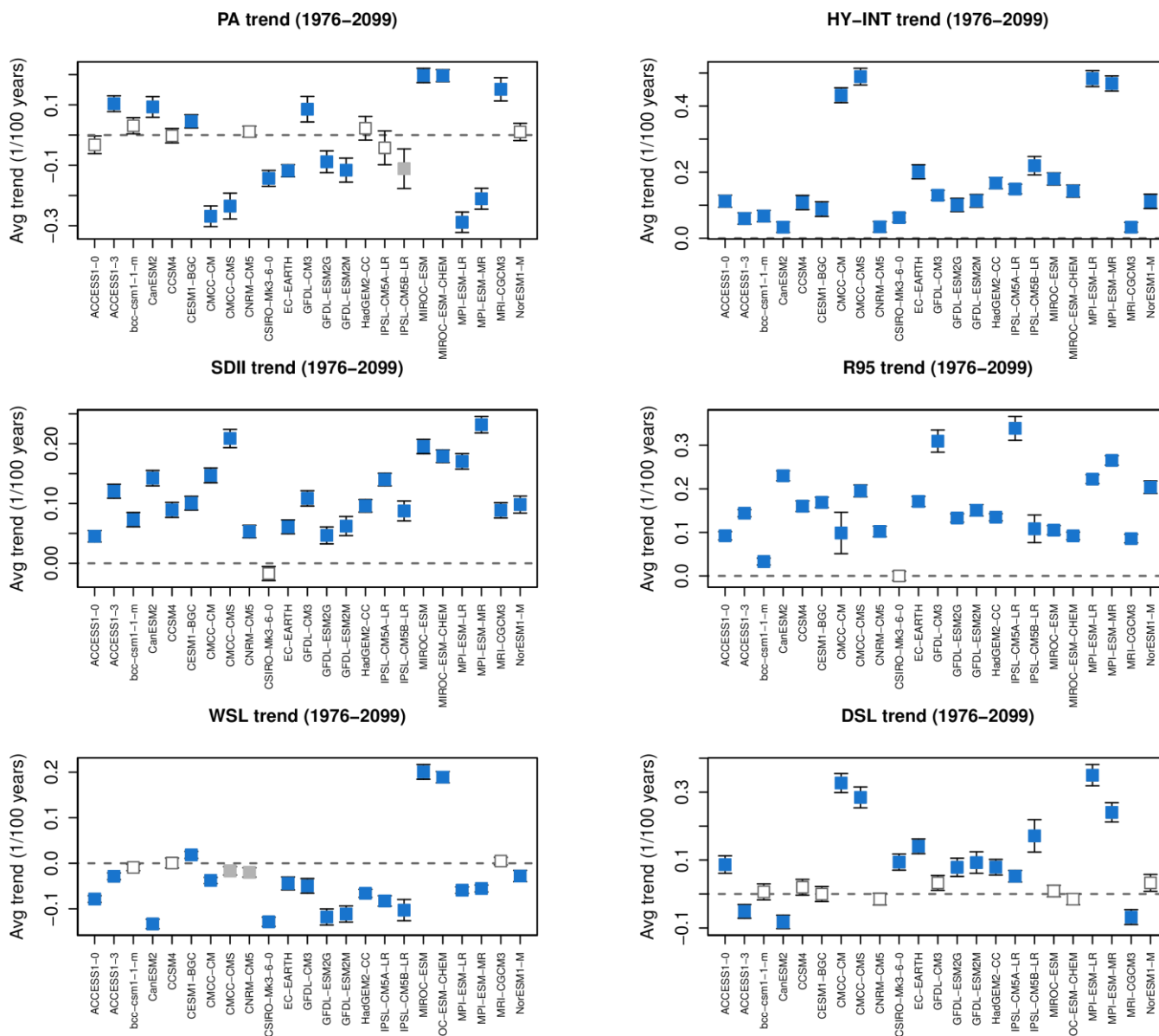


Figure 2: Trend in selected indices for an ensemble of CMIP5 models (historical + RCP 8.5 projection) over the time period 2006–2099. The trends are calculated over the latitude band 60°S–60°N. Data were normalized to the historical 1976–2005 period. Indices include the precipitation area (PA), hydroclimatic intensity (HY-INT), precipitation intensity (SDII), heavy precipitation (R95), wet and dry spell length (WSL and DSL) following Giorgi et al. 2014. Error bars show the geographical variability (standard deviation) within the region, colours the statistical significance of the trend (90% grey, 95% blue). This is an example of a large number of different plots which can be produced with *recipe_hyint.yml*, similar to Giorgi et al. (2014). For details see Section 3.1.1.

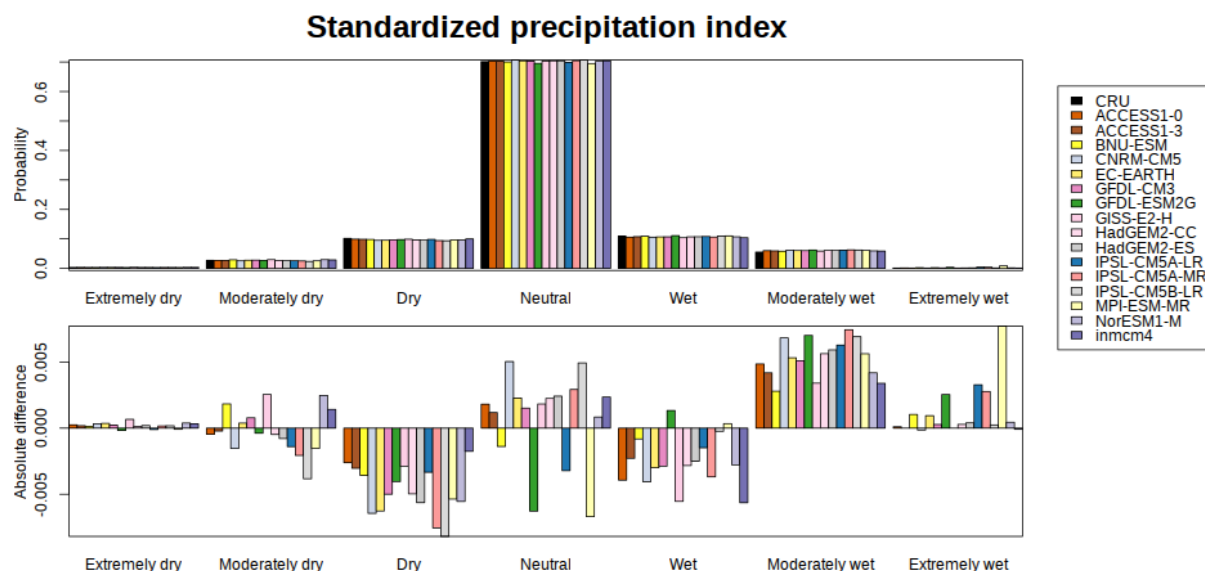


Figure 3: Output from SPI diagnostic in *recipe_spei.yml* with globally averaged histograms of SPI over land areas, weighted by the cosine of latitude for a selection of CMIP5 models and using gridded observations from CRUs4.01. (top) Absolute values, and
 695 (bottom) bias of all models compared to CRUs4.01, for details see Section 3.1.2.

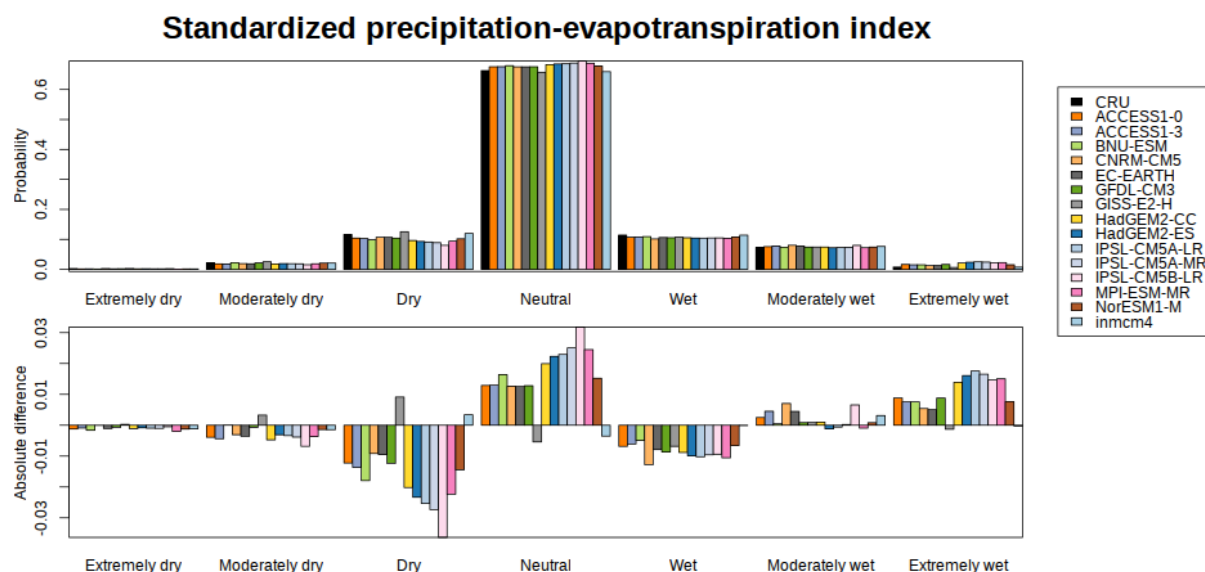


Figure 4: Output from the SPEI diagnostic in *recipe_spei.yml* with globally averaged histograms of SPEI over land areas, weighted



by the cosine of latitude for a selection of CMIP5 models and using gridded observations from CRUts4.01. (top) Absolute values, and (bottom) bias of all models compared to CRUts4.01, for details see Section 3.1.2.

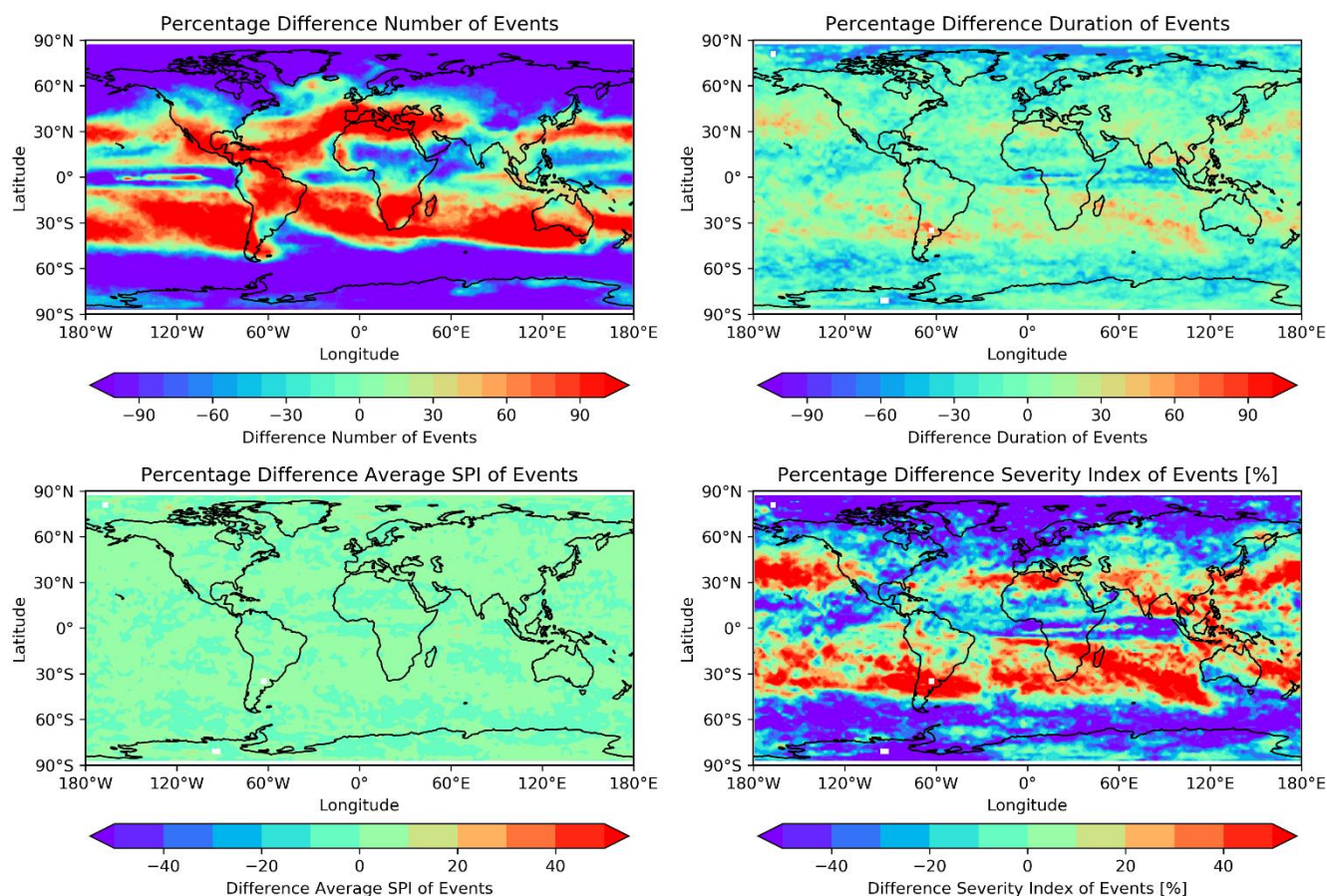


Figure 5: Difference in number (top left), duration (top right), average SPI (bottom left), and severity index (bottom right) of drought events between the RCP8.5 (2050-2100) and historic (1950 to 2100) multi-model mean of 13 CMIP5 models. Here, a drought event is defined as any number of consecutive months with an SPI < -2. For the SPI calculation a Gamma distribution and a representative time scale of 6 months is used. The Figure is similar to Figure 3 a – d of (Martin, 2018) and produced with recipe_martin18grl.yml, for details see Section 3.1.2.

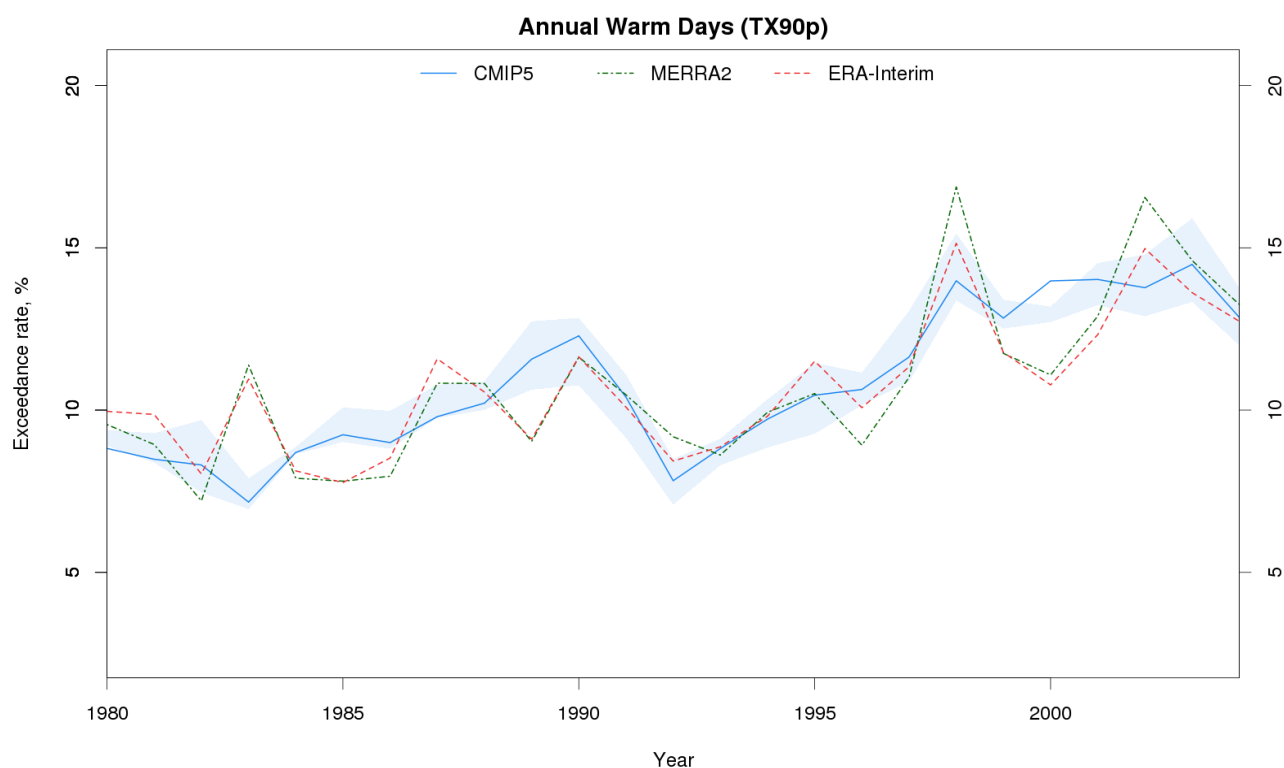


Figure 6: Time series plot of the annual percentage of days when the daily maximum temperature is higher than the 90th percentile for the respective calendar day. Percentile thresholds are calculated following (Zhang et al., 2005b) for the base period 1981–2010. The shading indicates the interquartile ensemble spread (range between the 25th and 75th quantiles). The CMIP5 ensemble mean (blue line, 5 models in this example) averaged over all land grid boxes is compared with the reanalysis datasets MERRA-2 (green dashed line) and ERA-Interim (red dashed line). Similar to Figure 9.37 e of IPCC AR5 (Flato et al., 2013) and produced with *recipe_extreme_events.yml*, for details see Section 3.2.

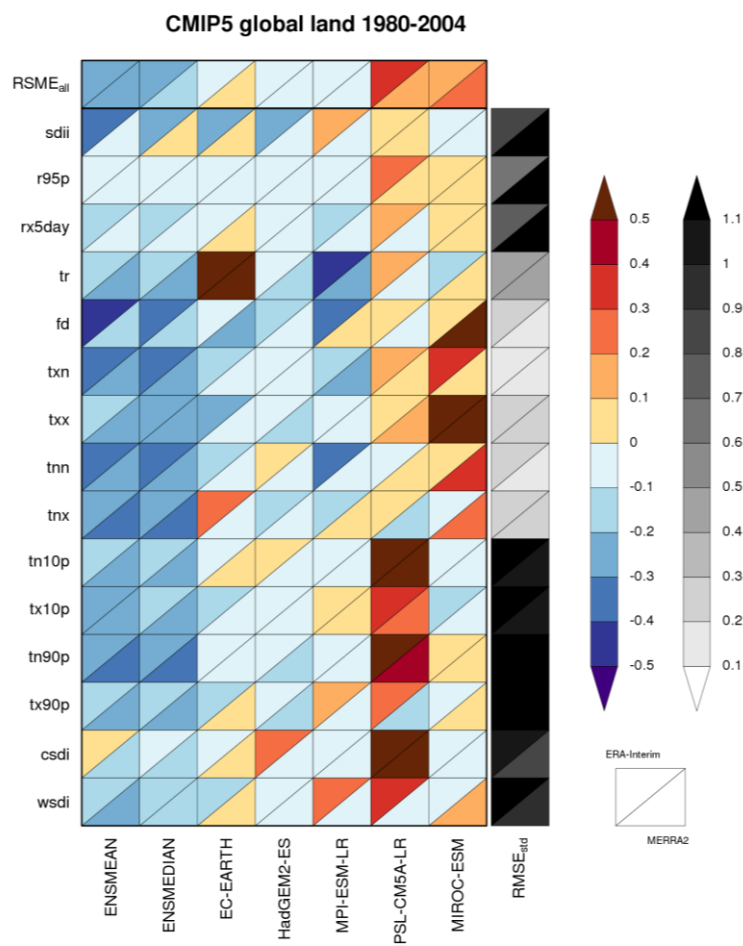


Figure 7: “Portrait” diagram showing relative spatially averaged root-mean-square errors (RMSE) in the 1981–2000 climatologies of 12 temperature and 3 precipitation indices simulated by CMIP5 models (5 in this example along x-axis) with respect to the two reanalyses ERA-Interim (upper triangle) and MERRA2 (lower triangle). The RMSEs are spatially averaged over all land grid points. The top row (RMSE_{all}) indicates the mean relative RMSE across all indices for the CMIP5 ensemble mean (first column) and median (second column) and each model individually. Blue (red) colours indicate that a model performs better (worse) than the ensemble mean error compared to the respective reanalysis dataset. The grey shaded column at the right-hand side indicates the median RMSE normalized by the spatial standard deviation of the index climatology in the reanalyses (RMSE_{std}). The root-mean-square error is shown in greyscale on the right. See (Sillmann et al., 2013a) for details. Similar to Figure 9.37a of the IPCC AR5 report (Flato et al., 2013) and produced with *recipe_extreme_events.yml*, for details see Section 3.2.

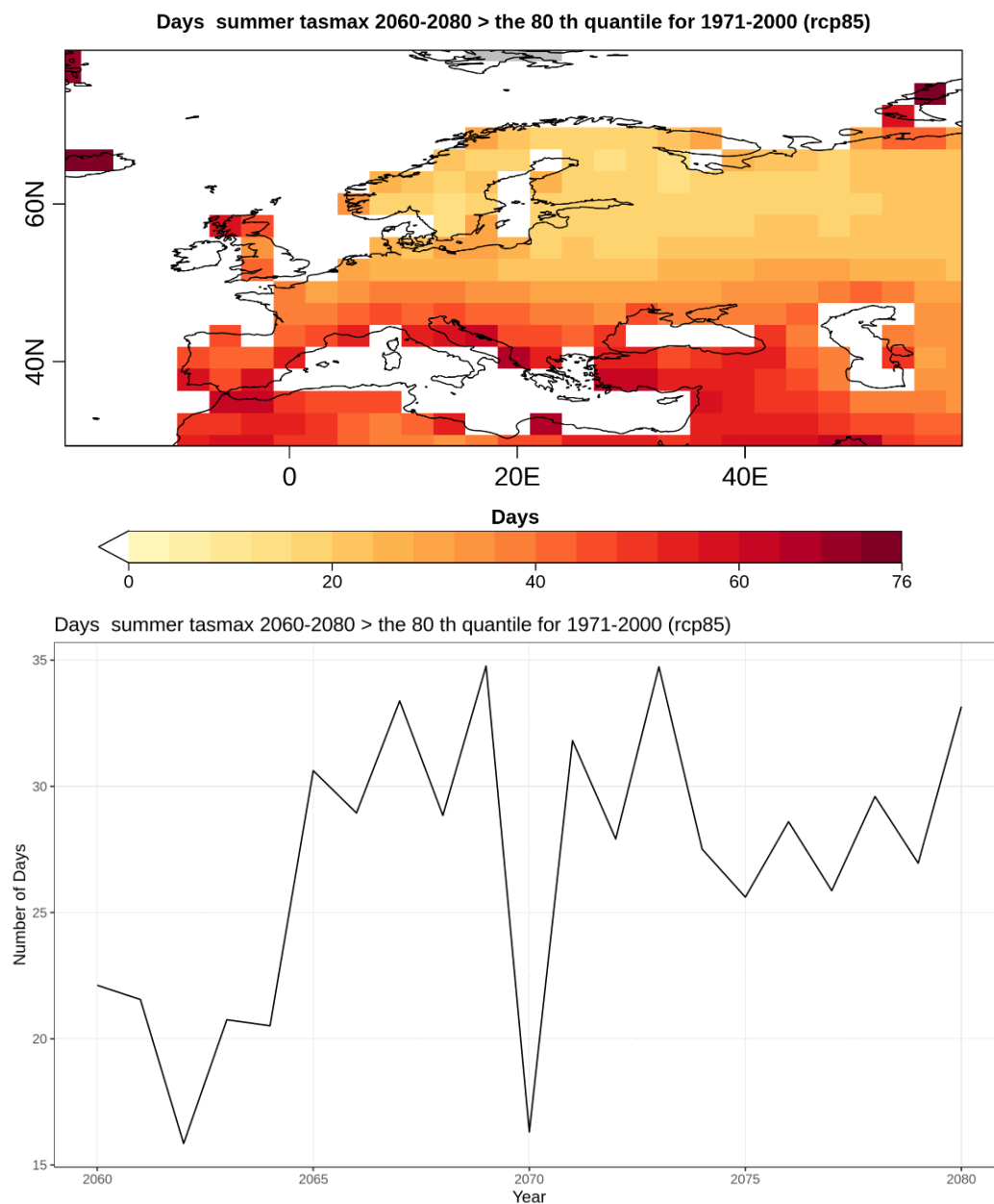


Figure 8: a) Average number of summer days during the time period 2060-2080 when the daily maximum near-surface air temperature exceeds the 80th quantile of the 1971-2000 reference period. b) Mean annual number of summer days when the daily maximum near-surface air temperature exceeds the 80th quantile of the 1971-2000 reference period averaged over the region shown in a). Results shown are for the RCP 8.5 scenario simulated by BCC-CSM1-1 (see Section 3.3.1 for details on *recipe_heatwaves_coldwaves.yml*).

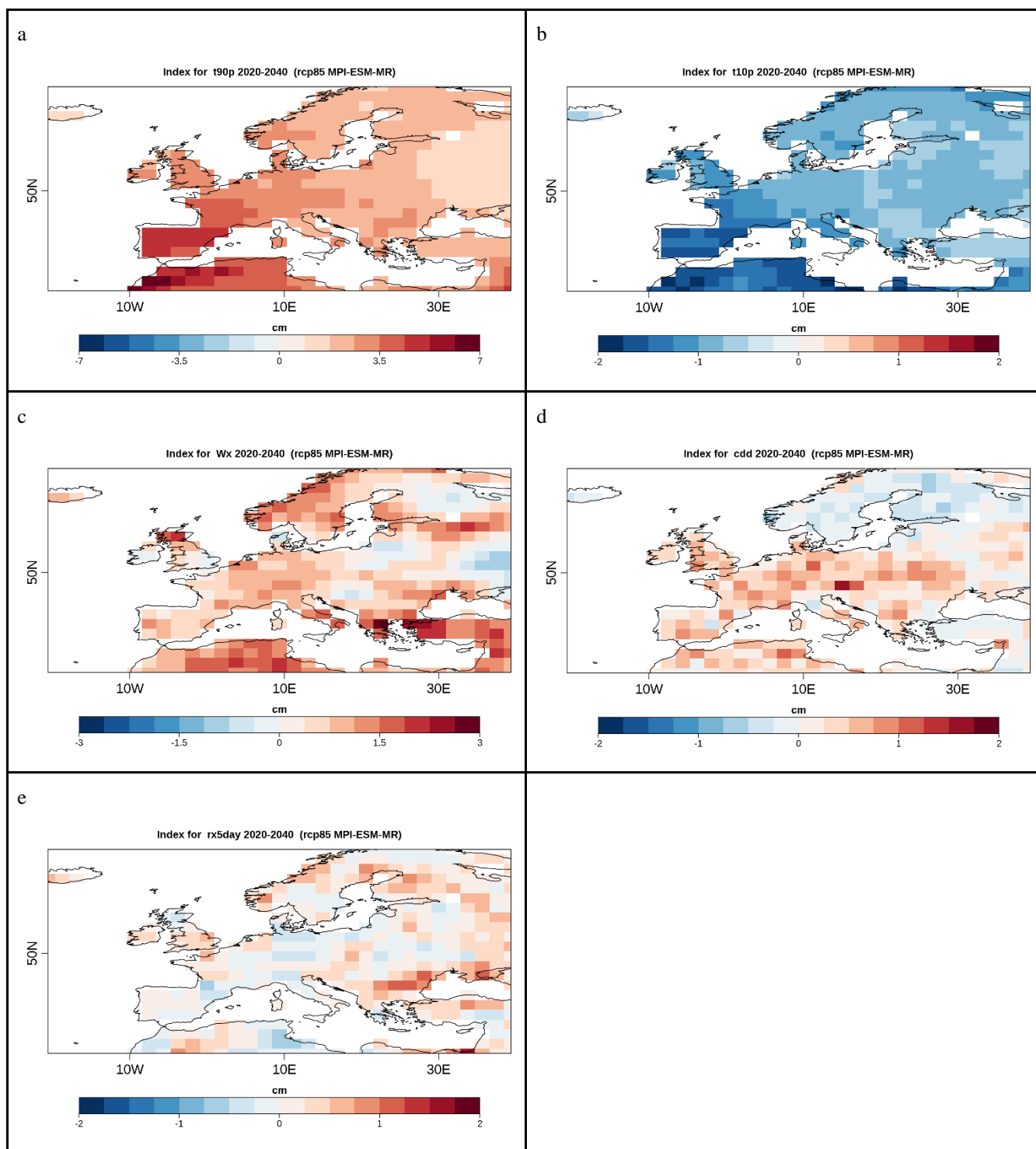


Figure 9: a-e) Average change in each of the components of the Combined Climate Extreme Index for the time period 2020-2040 compared to the 1971-2000 reference period (a) upper temperature percentile, b) lower temperature percentile, c) wind, d) precipitation, e) drought).The results are shown for the RCP 8.5 scenario simulated by MPI-ESM-MR (see Section 3.3.2 for details on *recipe_extreme_index.yml*).

Number of days exceeding the DTR by 5 degrees
during the period 2030-2080

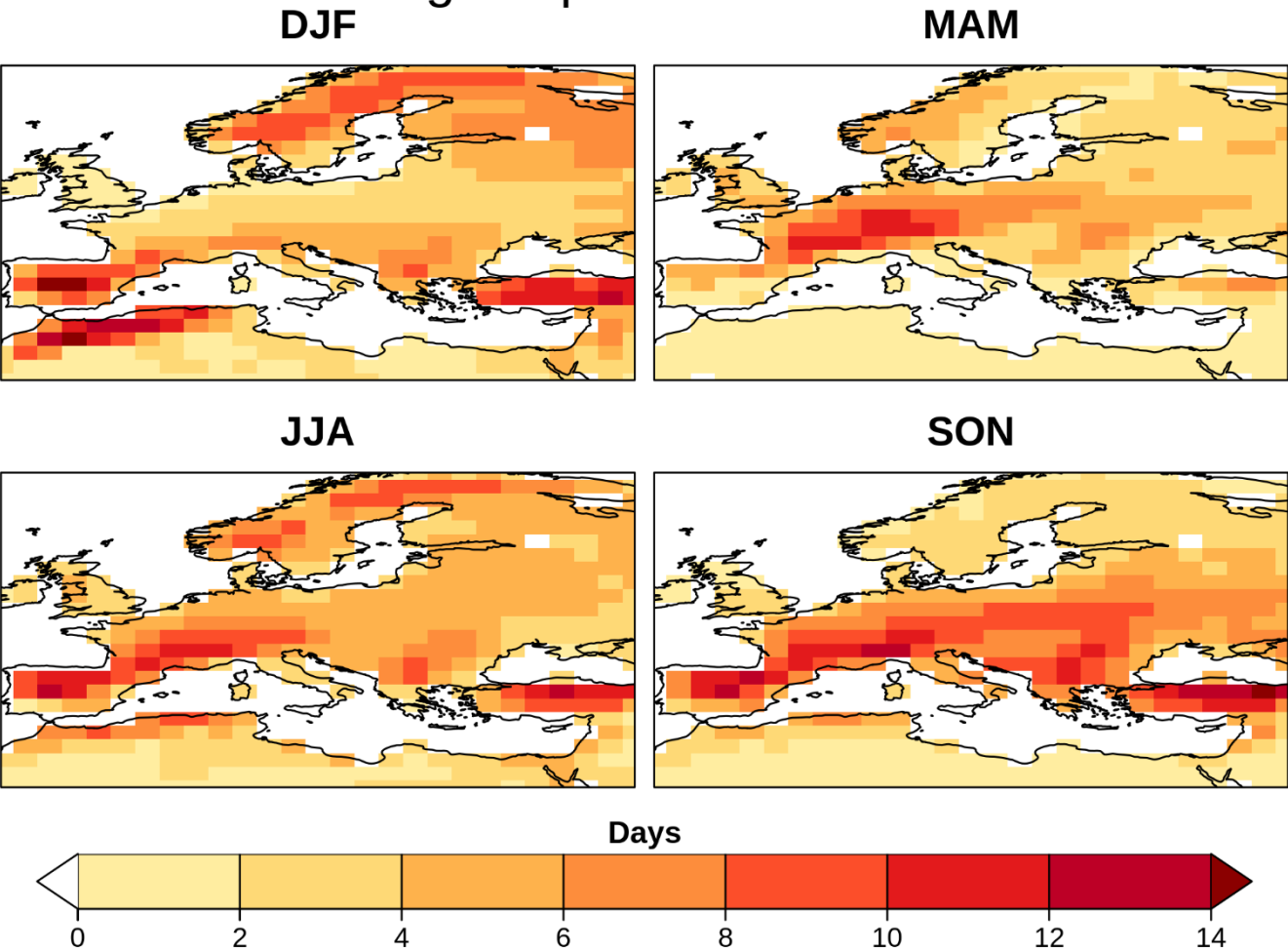


Figure 10: Average number of days/year exceeding the Diurnal Temperature Range (DTR) of the historical period (1961-1990) by 5 degrees during the period 2030-2080. The example shown is calculated for the RCP 8.5 scenario simulated by MPI-ESM-MR (see Section 3.3.3 for details on *recipe_diurnal_temperature_index.yml*).



CF from IPSL-CM5A-MR (2021-2050)

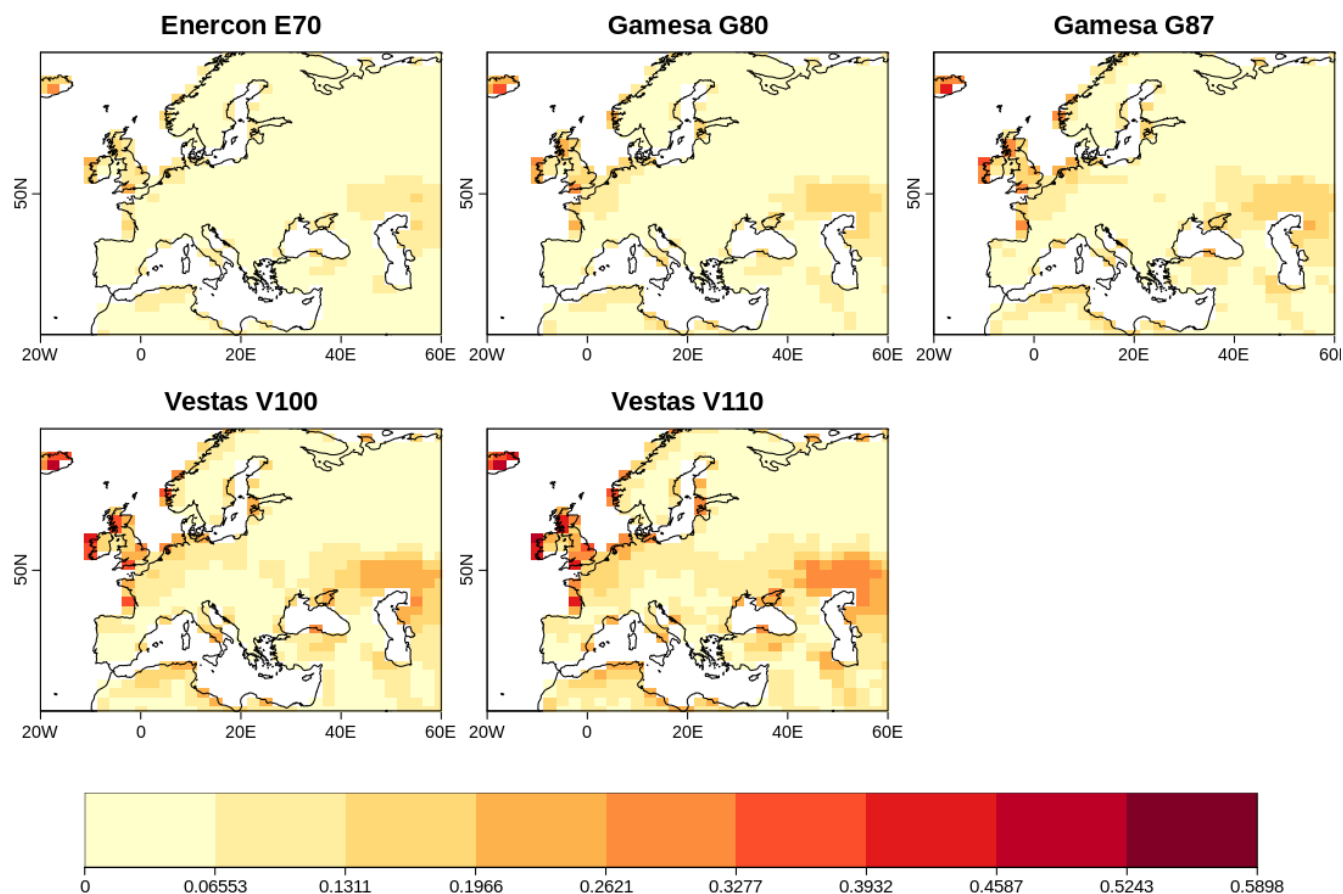


Figure 11: Wind capacity factor for five kinds of wind turbines: Enercon E70 (top-left), Gamesa G80 (middle-top), Gamesa G87 (top-right), Vestas V100 (bottom-left) and Vestas V110 (middle-bottom) using the IPSL-CM5A-MR simulation for the rcp8.5 scenario during the period 2021-2050 (see Section 3.3.4 for details on *recipe_capacity_factor.yml*).

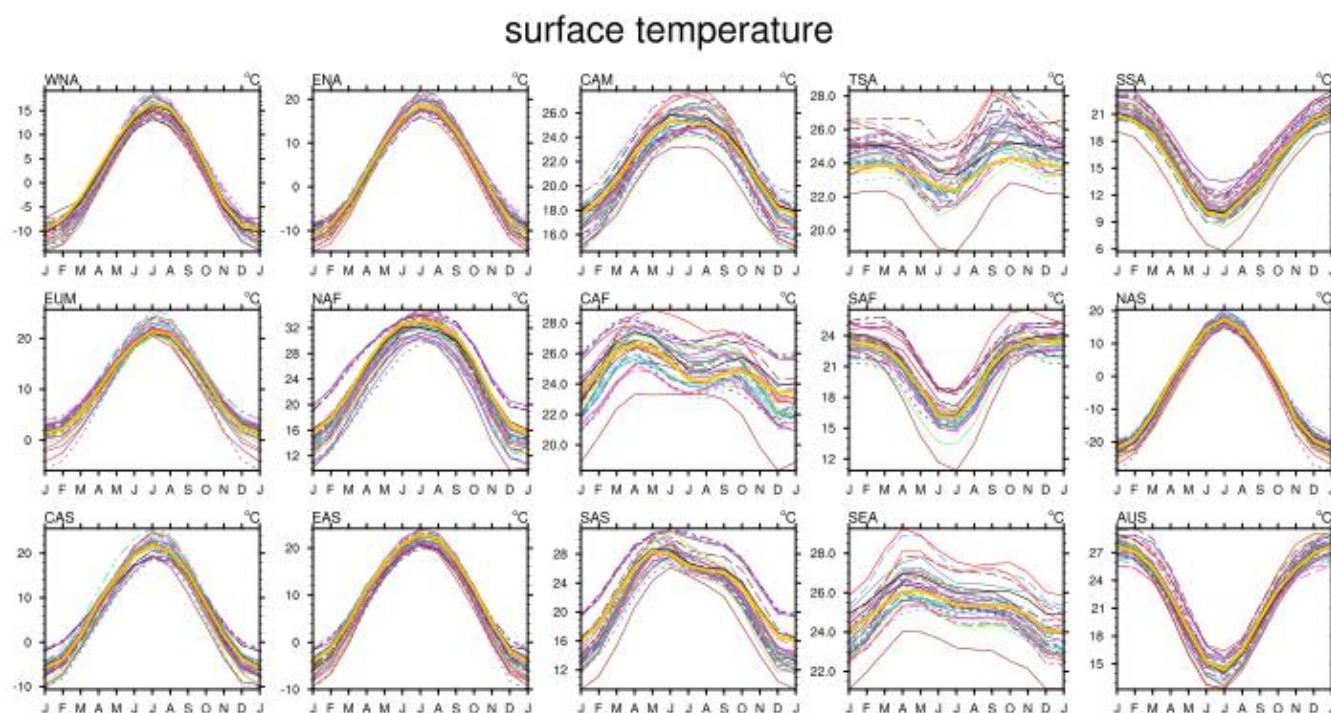
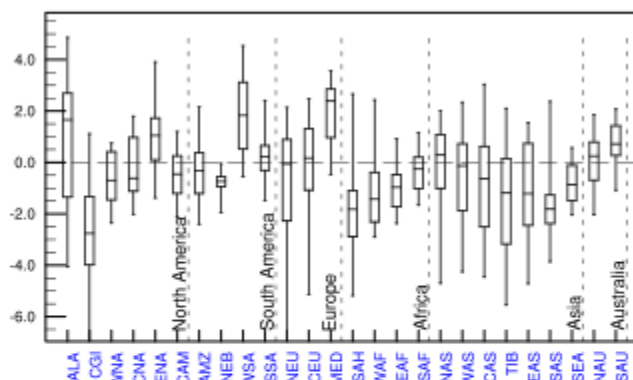


Figure 12: Mean seasonal cycle for the surface temperature (tas) averaged for 1980-1999 over land in different regions: Western North America (WNA), Eastern North America (ENA), Central America (CAM), Tropical South America (TSA), Southern South America (SSA), Europe and Mediterranean (EUM), North Africa (NAF), Central Africa (CAF), South Africa (SAF), North Asia (NAS), Central Asia (CAS), East Asia (EAS), South Asia (SAS), Southeast Asia (SEA), and Australia (AUS). There are 35 CMIP5 models as well as ERA-Interim and CRU data shown. Similar to Figure 9.38a of the IPCC AR5 report (Flato et al., 2013) and produced with *recipe_flato13ipcc.yml*, for details see Section 3.4.1.

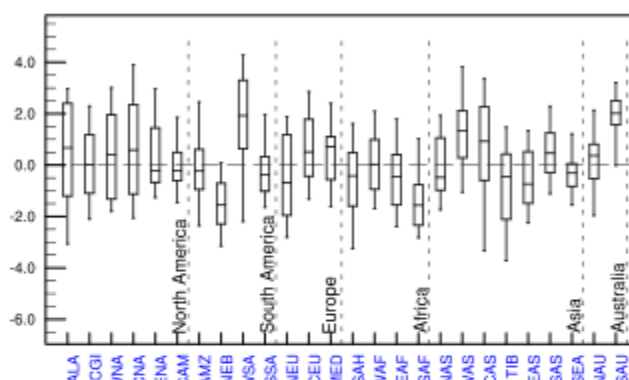


surface temperature bias vs. CRU

bias, surface temperature ($^{\circ}\text{C}$), DJF



bias, surface temperature ($^{\circ}\text{C}$), JJA



bias, surface temperature ($^{\circ}\text{C}$), ANN

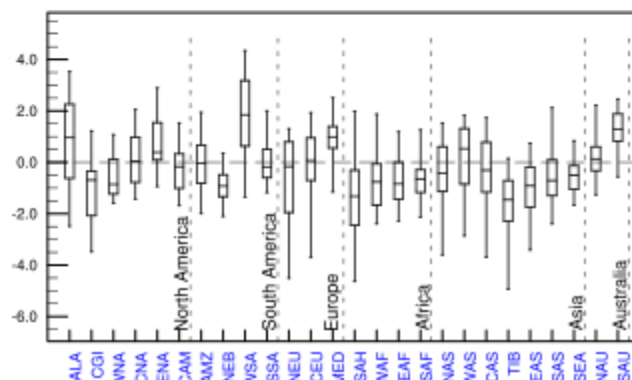


Figure 13 Box and whisker plots showing the 5th, 25th, 50th, 75th and 95th percentiles of the seasonal- and annual mean biases for the surface temperature (tas) between 34 CMIP5 models and CRU data. The regions are: Alaska/NW Canada (ALA), Eastern Canada/Greenland/Iceland (CGI), Western North America(WNA), Central North America (CNA), Eastern North America (ENA), Central America/Mexico (CAM), Amazon (AMZ), NE Brazil (NEB), West Coast South America (WSA), South-Eastern South America (SSA), Northern Europe (NEU), Central Europe (CEU), Southern Europe/the Mediterranean (MED), Sahara (SAH), Western Africa (WAF), Eastern Africa (EAF), Southern Africa (SAF), Northern Asia (NAS), Western Asia (WAS), Central Asia (CAS), Tibetan Plateau (TIB), Eastern Asia (EAS), Southern Asia (SAS), Southeast Asia (SEA), Northern Australia (NAS) and Southern Australia/New Zealand (SAU). Similar to Figure 9.39a,c,e of the IPCC AR5 report (Flato et al., 2013) and produced with *recipe_flato13ipcc.yml*, for details see Section 3.4.1.



Precipitation Rate bias vs. CRU

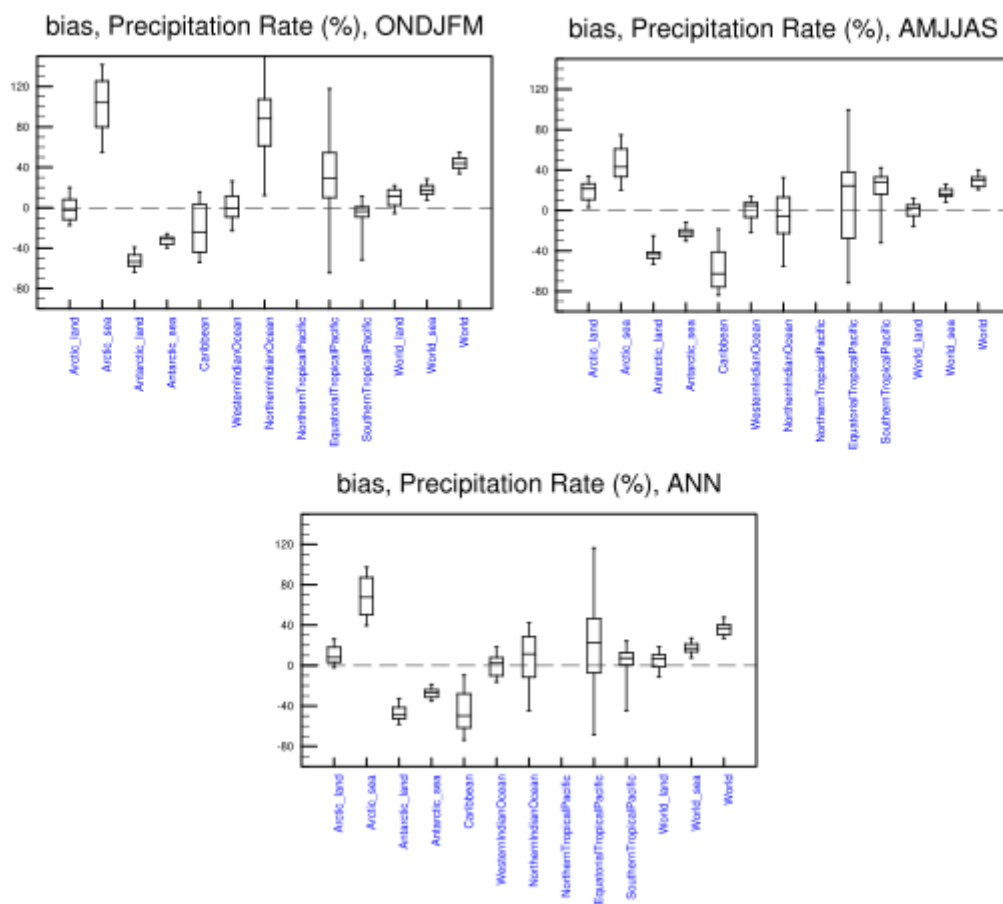


Figure 14: Box and whisker plots showing the 5th, 25th, 50th, 75th and 95th percentiles of the seasonal- and annual mean biases for the precipitation (pr) for oceanic and polar regions between 34 CMIP5 models and CRU data. Similar to Figure 9.40b,d,f of the IPCC AR5 report (Flato et al., 2013) and produced with *recipe_flato13ipcc.yml*, for details see Section 3.4.1.



Centred surface temperature bias vs. ERA-Interim

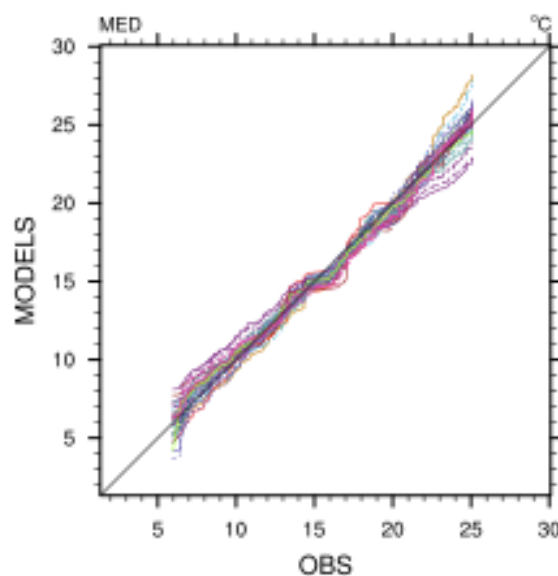


Figure 15: Ranked modelled versus ERA-Interim mean temperature for 34 CMIP5 models in the Mediterranean region for the 1979–2000 period. Similar to Figure 9.41b of the IPCC AR5 report (Flato et al., 2013) and produced with *recipe_flato13ipcc.yml*, for details see Section 3.4.1.

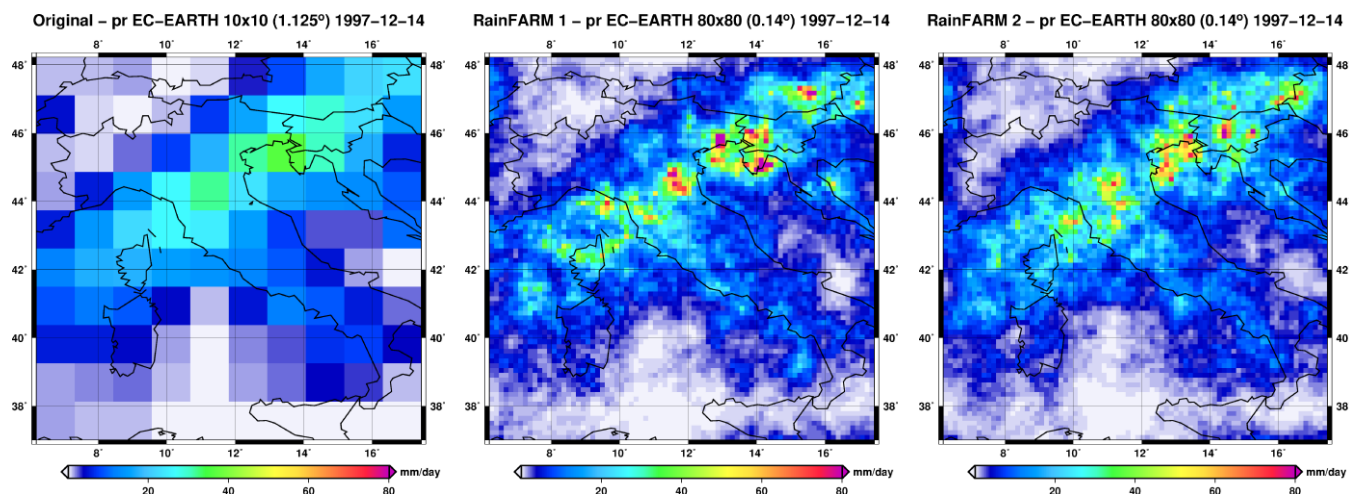


Figure 16: Left panel: Example of daily accumulated precipitation from the EC-EARTH CMIP5 model on a specific day, downscaled using RainFARM from its original resolution (1.125°). Central and right panel: Two stochastic realizations for increasing the spatial resolution by a factor of 8 to 0.14° ; A fixed spectral slope of $s=1.7$ was used. The data were produced by *recipe_rainfarm.yml* but this plot was not produced by ESMValTool - the recipe output is netCDF only.

