Copernicus Climate Change Service

Validation framework

Issued by: CSIC-IFCA

Date: 30/06/2021

Ref: C3S_D34d.3.1.1_validation_v4

Official reference number service contract: 2019/C3S_34d_CSIC-IFCA/SC1

This document has been produced in the context of the Copernicus Climate Change Service (C3S). The activities leading to these results have been contracted by the European Centre for Medium-Range Weather Forecasts, operator of C3S on behalf of the European Union (Delegation Agreement signed on 11/11/2014). All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission and the European Centre for Medium-Range Weather Forecasts has no liability in respect of this document, which is merely representing the authors view.

IMPLEMENTED BY ECMWF

Contributors

CSIC-IFCA

Javier Díez
Sixto Herrera
Antonio S. Cofiño
José M. Gutiérrez

SMHI

Grigory Nikulin

Predictia

Daniel San Martín
Max Tuni

Table of Contents

1. Introduction 4

2. Evaluation framework 6

2.1. Ensemble diagnostics 8

2.2. Evaluation indices and metrics 9

2.3. Observational uncertainty 12

3. Summary 12

4. References 13

1. Introduction

There is a vast literature on the evaluation of Regional Climate Models (RCMs) over particular CORDEX domains[1]. Most of these studies build on reanalysis-driven (ERA-Interim) RCM simulations (*validation runs*) and analyze the intrinsic errors of the RCMs driven by "perfect" boundary conditions (e.g. cold/warm or wet/dry biases) to assess their suitability for the region of interest. Fewer studies analyze RCM simulations driven by Global Circulation Models (GCMs) using *historical runs* to assess the improvement or added value of the regional model (e.g. smaller biases or better spatial representation than the global model outputs) and to inform future projections produced with the same GCM/RCM pairs under different emission scenarios. These studies have proposed a large number of evaluation indices to assess different aspects of the regional climate, including distributional statistics (central tendency, variability, and extremes), indices characterizing the temporal (spells, annual cycle, inter-annual variability and trends) or spatial (covariance) structure, and indices for inter-variable dependency.

The initiative VALUE (Validating and Integrating Downscaling Methods for Climate Change Research; Maraun et al. 2015) conducted a categorization and a selection of a reduced set of common validation indices to provide synthesized evaluation information for a number of aspects relevant for different user applications: distributional, temporal, spatial and inter-variable. These categories include different indices characterizing the mean and extreme regimes, thus informing for different dimensions of analysis which are relevant for practical applications (e.g. extreme spells are relevant in agriculture, whereas extreme distributional values are critical in health applications).

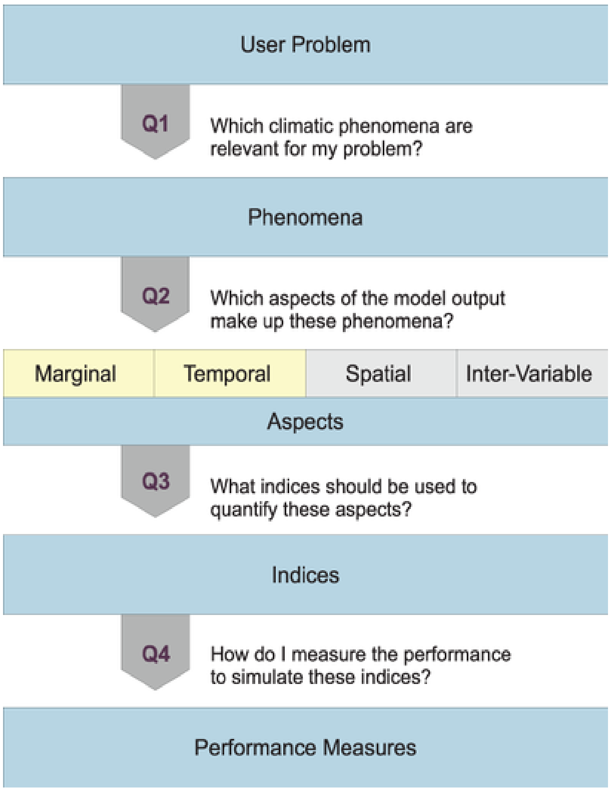
User Problem
Q1
Which climatic phenomena are relevant for my problem?
Phenomena
Q2
Which aspects of the model output make up these phenomena?
Marginal
Temporal
Spatial
Inter-Variable
Aspects
Q3
What indices should be used to quantify these aspects?
Indices
Q4
How do I measure the performance to simulate these indices?
Performance Measures

Figure 1. Validation tree (Maraun et al. 2015).

[1] https://cordex.org/publications/peer-reviewed-publications/

In this document we present a simple evaluation framework which is designed to provide key summary information for the users of the C3S worldwide CODEX dataset uniformly across all regions. The goal is not providing detailed model by model information, but producing summary information of the full ensembles available for the different domains, so the users can quickly analyze the homogeneity of the dataset for the different variables and aspects of interest and get some preliminary information to understand the performance of the different models relative to the full ensemble. This information could be also useful to inform sub-ensemble selection for those applications where a reduced number of models is required (e.g. to feed impact models).

To this aim, we followed the approach introduced in VALUE and selected a reduced number of indices providing basic information for the different dimensions (mean values and extremes) for two aspects of interest (distributional and temporal aspects). The proposed selection is aligned with some recent validation studies produced by the CORDEX-CORE community[2] across several domains (Teichmann et al. 2020, Coppola et al., 2021) and also with the preliminary validation results included in the IPCC-AR6 Atlas (the C3S worldwide CORDEX evaluation is designed as an extension of the basic validation results included in the IPCC report). The indices commonly used to characterize spatial and inter-variable structure are typically multi-dimensional and difficult to compute and summarize across regions (Widmann et al. 2019). Therefore, we have focused on the mean and extreme dimensions of distributional and temporal aspects in this proposal in order to keep it realistic and informative for different user's needs.

The CORDEX worldwide C3S dataset to be evaluated is described in D34d.1.1.3 (final inventory). It presents the simulations forming the ensembles of the different domains and describes the technical evaluation and curation (metadata fixing) performed on this data, as compared to the original information available in the Earth System Grid Federation (ESGF). Here we focus on the scientific validation which, jointly with the technical information in D34d.1.13, provides comprehensive evaluation of the C3S worldwide CORDEX dataset on both technical (metadata) and scientific (values) aspects relevant for end users.

[2] CORDEX-CORE is a CORDEX activity designed to produce homogeneous regional projections across nine CORDEX domains at 0.22° resolution (covering major inhabited regions); https://cordex.org/experiment-guidelines/cordex-core/cordex-core-simulations/

For particular region- or process-dependent diagnostics and evaluation indices the user is referred to the selection of papers for the different CORDEX domains provided in the CDS documentation[3], or to the complete list of papers compiled in the CORDEX web[4].

2. Evaluation framework

The main goal of the proposed evaluation framework is providing homogeneous information across the different domains –thus favoring comparability– and consists of two phases:

1. Computing ensemble **diagnostics** (statistics obtained directly from model data) providing a visual outlook suitable to identify strange values and/or outliers in the data and/or models that deviate from the ensemble.
2. Computing **evaluation indices** (statistics obtained comparing model data and reference observations) providing objective measures on model performance informing on the merits and limitations of the simulations forming the ensemble.

Both diagnostics and evaluation indices can be computed from the available *evaluation* (ERA-Interim-driven) and *historical* (GCM-driven) simulations (see final inventory in D34d.1.1.3), providing complementary information for end users: 1) intrinsic performance of the RCM and 2) conditioned performance when driven by a particular GCM. Therefore, we will consider both experiments in this framework and calculate diagnostics and evaluation indices for both, as well as for the driving models (ERA-Interim or the particular GCM).

The C3S worldwide CORDEX dataset covers sixteen common variables. The six core ones are shown in Table 1 and will form the focus of the evaluation framework. Note that most RCM evaluation studies focus on temperature and/or precipitation, with little information on other variables. Diagnostics will be computed for the six variables, providing some basic information of the ensemble composition and allowing end users to visualize the available ensembles (e.g. using climate stripes, which provide visual synthesis information of ensembles; see Sec. 2.1). Evaluation indices will be computed for the three selected variables: precipitation, temperature and wind.

[3] https://confluence.ecmwf.int/display/CKB/CORDEX%3A+Regional+climate+projections

[4] https://cordex.org/publications/peer-reviewed-publications/

Number	Variable	Code	Units
1	Precipitation	pr	kg m-2 s-1
2	Mean surface air temperature	tas	m s-1
3	Near-surface wind speed	sfcWind	K
4	Maximum surface air temperature	tasmax	K
5	Minimum surface air temperature	tasmin	K
6	Near-surface specific humidity	huss	1

Table 1. Six surface climate variables considered in the evaluation framework. All six variables will be used to produce ensemble diagnostics, whereas evaluation results will be produced for the first three.

Results are typically displayed as maps (with gridbox results of the diagnostic or index) and also averaged over regions in order to provide summary user-friendly information in the form of tables or color matrices facilitating a quick overview of the results. The regions considered depend on the spatial scale of the study and range from continental or subcontinental to national scale. For instance, the IPCC regions (Iturbide et al. 2020) are widely used for global studies since they represent homogeneous climatic regions worldwide in terms of climatological conditions and future climate projections. These regions have been used in the recent studies evaluating CORDEX CORE results across domains (Teichmann et al. 2020, Coppola et al., 2021). Smaller regions (e.g. the PRUDENCE regions for Europe) have been used in domain (or sub-domain) specific validation studies, such as Vautard (2020), but these have too much granularity for worldwide studies. Therefore, in this framework we will use the updated IPCC reference regions (see Figure below) for providing regional RCM validation results. This has the added value of aligning with CORDEX evaluation results included in AR6-WG1, which use these regions as reference.

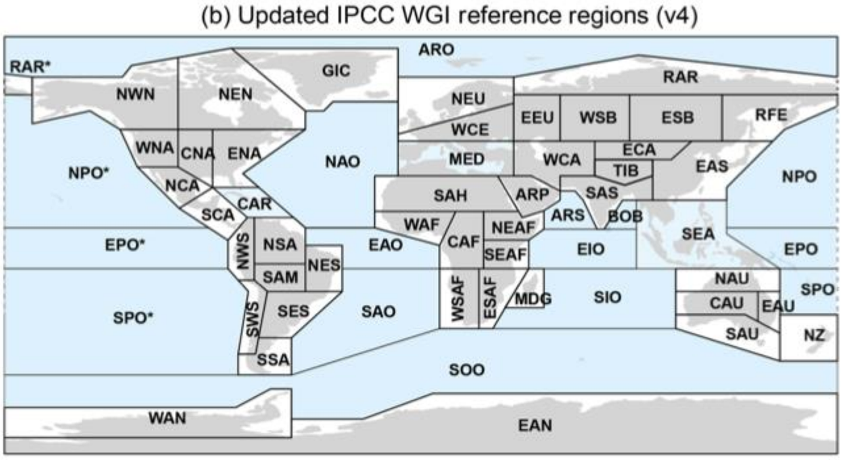
(b) Updated IPCC WGI reference regions (v4)

Figure 2. Updated IPCC reference regions showing 46 land (grey shading) and 15 ocean (blue shading) regions; stars indicate regions going across the zero meridian which are defined in two parts.

2.1. Ensemble diagnostics

Although diagnostics provide and outlook of the ensemble and allow for quick outlier or model departure detection, their use is scarce in the RCM validation literature. Recently, Vautard (2020) have used diagnostics (in particular matrix plots of climatological mean values) to represent the seasonal values of the large ensemble of EUR-11 simulations. Here we propose the use of climate stripes[5] showing visually the time series of annual values (in columns) for the different available RCM simulations (in rows). Figure 3 shows an example of the proposed approach for a representative domain (CORDEX-NAM), variable (mean temperature), and scenario (historical).

Figure 3 efficiently conveys different pieces of information relevant for end users. For instance, most of the models don't provide data in the first year (1950) of the historical period and three of them start in 1971. The plots also show the different ranges of annual values and interannual variability across models (in rows), thus providing a visual summary of the ensemble and allowing to identify abnormal simulations (large differences relative to the rest of the ensemble members) and/or values for specific years.

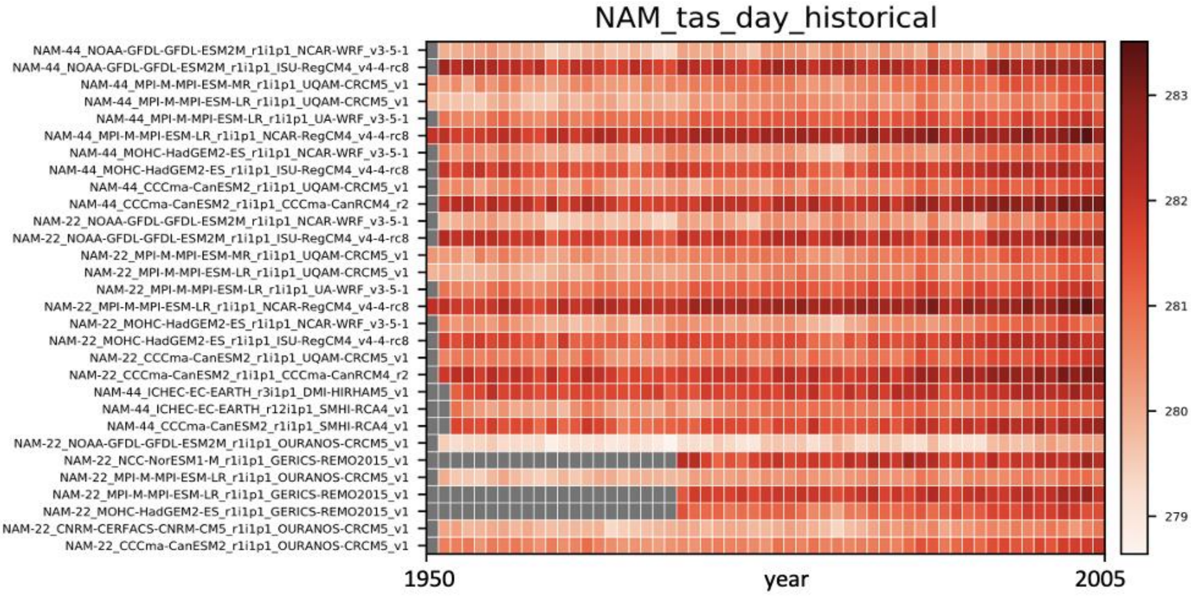

Figure 3. Example of climate stripes for CORDEX-NAM ensemble aggregated showing the domain-averaged annual values (in columns) of the different simulations forming the CORDEX-NAM historical ensemble (in rows). Gray denotes lack of data.

[5] Climate stripes are a new graphical diagnostic consisting of a series of coloured stripes chronologically ordered to visually portray climatological values and variability.

2.2. Evaluation indices and metrics

We build on the work done in the VALUE initiative to provide suitable indices characterizing various aspects/dimensions which are relevant for practical applications. The fit-for-purpose validation indices initially proposed in VALUE are shown in Table 2 indicating whether the indices analyze distributional, temporal and/or extreme aspects. Validation ultimately consists of deriving specific indices characterizing those aspects (e.g. mean value, dry spells, or 98th percentile) for both model output and observations (WFDE5, bias adjusted ERA5 is used as the reference observational dataset in this work; Cucchi et al., 2020) and quantifying the possible mismatch between them with the help of suitable performance measures (bias, relative bias -or ratio-, etc.). All the VALUE indices and measures have been implemented in R and are collected in the package VALUE[6], allowing for further extension, as well as for research reproducibility.

Index Code	**Measure**	**P**	**O**	**Description**	**C**	**T**	**E**
Mean	**Bias (% for P)**	✓	✓	**Mean value**	✓		
Variance	**Bias (%)**	✓	✓	**Quasi-variance**	✓		
Upper-MaxSpell	**bias**		✓	**Median of the upper-tail spell maxima (maximum number of consecutive days with value > 90th percentile)**		✓	✓
Lower-MaxSpell	bias		✓	Median of the annual lower-tail spell maxima (maximum number of consecutive days with value < 90th percentile)		✓	✓
UpperSpellP50	bias		✓	Median of the upper-tail spells (maximum number of consecutive days with value > 90th percentile)		✓	
LowerSpellP50	bias		✓	Median of the lower-tail spells (maximum number of consecutive days with value < 10th percentile)		✓	
R01	relbias	✓		Wet-day frequency (number of days with precipitation >= 1mm)	✓		
SDII	relbias	✓		Mean wet-day precipitation	✓		
DryAnnual-MaxSpell	**bias (days)**	✓		**Median of the annual dry spell maxima (maximum number of consecutive days with precipitation < 1mm)**		✓	✓
WetAnnual-MaxSpell	bias (days)	✓		Maximum of the annual wet spell maxima (maximum number of consecutive days with precipitation >=1mm)		✓	✓

[6] https://github.com/SantanderMetGroup/VALUE

DWProb	bias (%)	✓		Dry-wet transition probability		✓	
WWProb	bias (%)	✓		Wet-wet transition probability		✓	
P98	**bias (%)**	✓	✓	**98th percentile (of precipitation amount on wet days for precip)**			✓
AnnualCycle-RelAmp	**bias (%)**	✓	✓	**Relative amplitude of the annual cycle (range of the monthly mean values divided by the annual mean)**	✓	✓	
Interannual-Var	**bias (%)**	✓	✓	**Interannual variability (std of the annual time series)**	✓	✓	

Table 2. Validation indices and measures used in VALUE. Index codes refer to the package VALUE (see Section 3). Second and third columns indicate whether the indices are applied for validating precipitation (**P**) or other (**O**, including temperature) variables, respectively. The last three columns indicate whether the indices analyze central tendency or variability (**C**), temporal (**T**) or extreme (**E**) aspects, respectively. Note that spells have been defined as at least two consecutive days fulfilling the particular condition. Adapted from Maraun et al. (2019), Table 1. Selected indices are bold faced.

More recently, Vautard et al. (2020) and Coppola et al. (2021) reported evaluation results for EURO-CORDEX and CORDEX-CORE simulations, respectively, using a subset of indices linked to some economic sectors (including some of the standard ETCCDI indices[7] for extremes), intending to provide a broad view of the ensemble's capacity to represent useful information for further use in science and decision making. These indices are also used in the AR6-WGI report to provide future information of extreme indices and climatic impact-drivers. Some of these indices are included in Table 3.

Index Code	**Measure**	**P**	**O**	**Description**	**C**	**T**	**E**
TGx	**bias**		✓	**Yearly maximum of mean daily temperatures**			✓
TGn	**bias**		✓	**Yearly minimum of mean daily temperatures**			✓
Rx1day	**bias (%)**	✓		**Highest 1-day precipitation amount**			✓
Rx5day	bias (%)	✓		Highest 5-day precipitation amount			✓
HDD	bias		✓	Heating degree days		✓	✓
CDD	bias		✓	Cooling degree days		✓	✓
SPI	bias	✓		Standardized Precipitation Index	✓	✓	
TX35	bias	✓		#days/year with maximum temperature > 35°C	✓		
SDII	Bias (%)	✓		Mean wet-day precipitation	✓		
FD	Bias (%)	✓		Frost days (minimum temperature ≤ 0°C)	✓		✓

Table 3. Indices and measures used in relevant CORDEX literature and also in the IPCC AR6-WGI report. Selected indices are bold faced.

[7] http://etccdi.pacificclimate.org/list_27_indices.shtml

In the present work we selected ten indices from Tables 2 and 3 which constitute a good representation of the distributional, temporal and extremes aspects for the different variables. These indices will be used to provide evaluation results consistently across the different CORDEX domains, extending the preliminary results which have been produced for the IPCC AR6 Atlas chapter.

Figure 3 shows a proposal of synthesis information for a particular evaluation index (bias of mean temperature) over the North America domain in RCM simulations driven by reanalysis and historical GCM simulations. Annual and seasonal (DJF and JJA) biases are computed for both the RCMs and driving GCMs. As mentioned before, biases in the reanalysis-driven RCMs result from intrinsic model errors, with the results displayed being spatially aggregated for each reference region. This same analysis is performed for the GCM-driven RCM simulations to allow comparison of the intrinsic bias of the RCMs with the biases resulting when driven by the different GCMs. This representation is used in the AR6 WGI report to provide some simple and preliminary evaluation of CORDEX results worldwide.
One of such figures per domain and variable would provide a manageable amount of information suitable to provide end users with compact information of the available ensembles across domains.

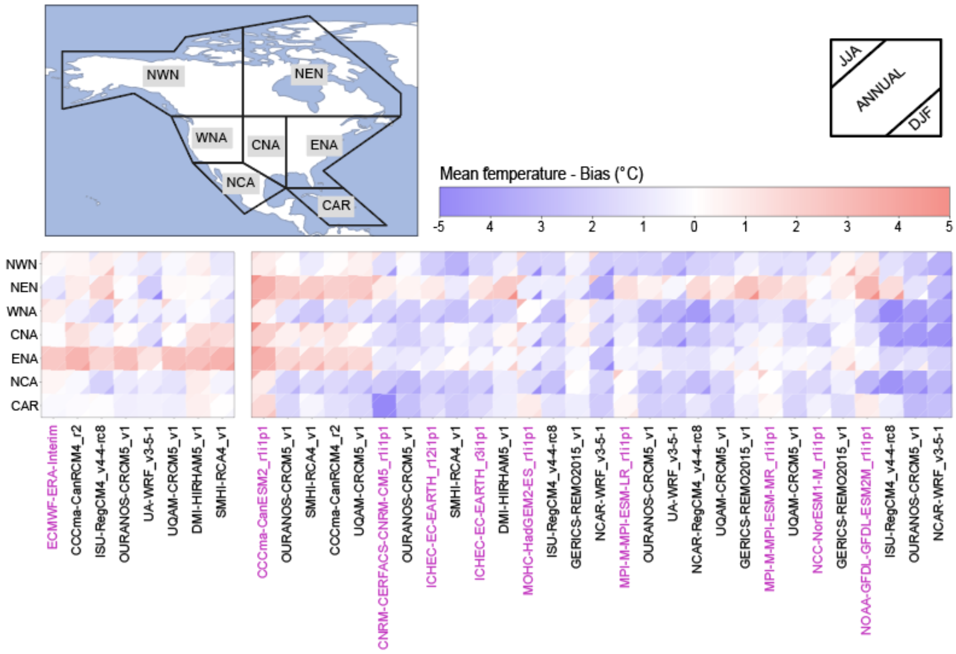

Figure 4. Example of the summary evaluation results for one particular evaluation index (mean temperature) a particular domain (CORDEX-NAM). The subdomains are (see Figure 2): NWN (Northwestern North America, NEN (Northeastern North America), WNA (Western North America), CNA (Central North America), ENA (Eastern North America), NCA (Northern Central America) and CAR (Caribbean).

2.3. Observational uncertainty

Several studies have quantified the influence of uncertainties in gridded observational reference data on regional RCM evaluation, showing that **observational uncertainty** partly translate into RCM evaluation uncertainty. For most cases observational uncertainty is smaller than RCM uncertainty, particularly if appropriate observational products are used (with high density of stations and gridboxes representing real values). Nevertheless, for individual sub-regions and performance metrics observational uncertainty can dominate, even in regions with good observational datasets such as Europe (Kotlarski et al. 2019). Overall, the existing literature indicates that observational uncertainty can be pronounced (particularly in worldwide applications) and needs to be taken into account. In the proposed evaluation framework we consider two observational datasets to provide evaluation results over those regions where observational uncertainty may result in significant differences (e.g. regions with complex topography), thus assessing the effect of this factor:

1) The **ERA5** (Hersbach et al., 2020) reanalysis outputs.
2) The ERA5 bias adjusted version **WFDE5** (Cucchi et al., 2020), with monthly values adjusted using CRU TS4.03 from CRU (Harris et al., 2020) for 1979 to 2018 for all variables and the GPCCv2018 full data product (Schneider et al., 2018) for rainfall rates for 1979 to 2016.

The WFDE5 **half-degree horizontal resolution** grid is used as the common grid to undertake the evaluation, and ERA5 and CORDEX outputs are regridded to this reference grid using conservative remapping. **Daily frequency** is used for all datasets, since most of the evaluation indices are computed from daily data. The evaluation **period considered is 1980-2005**, which is a compromise between having a long reference period and using the common period of all datasets.

3. Summary

We present a simple evaluation framework which is designed to provide key user-friendly evaluation information uniformly across regions for the users of the C3S worldwide CORDEX data. The goal is producing summary information of the full ensembles available in the different domains, so the users can quickly analyze the homogeneity of the dataset for the different variables and aspects of interest and get some preliminary information for sub-ensemble selection (for those applications where a reduced number of models is required, e.g. to feed impact models).

The proposal includes six variables and ten evaluation indices and, for each CORDEX domain and variable, proposes one figure for representing key diagnostics of the ensemble and one figure representing the results of the validation metrics. This task will produce 1) reusable open software implementing the framework, 2) a dataset with the evaluation indices, and 3) documents suitable to be included as supporting documentation for the datasets in the CDS (e.g. in the CDS documentation tab).

4. References

Coppola, E. et al., 2021: Climate hazard indices projections based on CORDEX-CORE, CMIP5 and CMIP6 ensemble. Climate Dynamics, doi:10.1007/s00382-021-05640-z.

Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Lange, S., Müller Schmied, H., Hersbach, H., and Buontempo, C.: WFDE5: bias-adjusted ERA5 reanalysis data for impact studies, Earth Syst. Sci. Data, 12, 2097–2120, https://doi.org/10.5194/essd-12-2097-2020, 2020.

Harris, I., Osborn, T. J., Jones, P., and Lister, D.: Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset, Sci. Data, 7, 109, https://doi.org/10.1038/s41597-020-0453-3, 2020.

Hersbach, H, Bell, B, Berrisford, P, et al. The ERA5 global reanalysis. Q J R Meteorol Soc. 2020; 146: 1999– 2049. https://doi.org/10.1002/qj.3803

Iturbide, M. et al. (2020) An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets. Earth Syst. Sci. Data, 12, 2959–2970, https://doi.org/10.5194/essd-12-2959-2020

Kotlarski, S, Szabó, P, Herrera, S, et al. (2019) Observational uncertainty and regional climate model evaluation: A pan-European perspective. Int J Climatol. 39: 3730– 3749. https://doi.org/10.1002/joc.5249

Maraun, D., Widmann, M., Gutiérrez, J.M., Kotlarski, S., Chandler, R.E., Hertig, E., Wibig, J., Huth, R. and Wilcke, R.A. (2015), VALUE: A framework to validate downscaling approaches for climate change studies. Earth's Future, 3: 1-14. https://doi.org/10.1002/2014EF000259

Maraun, D, Widmann, M, Gutiérrez, JM. (2019) Statistical downscaling skill under present climate conditions: A synthesis of the VALUE perfect predictor experiment. Int J Climatol, 39: 3692– 3703. https://doi.org/10.1002/joc.5877

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M., and Rudolf, B.: GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle, Theor. Appl. Climatol., 115, 15–40, https://doi.org/10.1007/s00704-013-0860-x, 2014.

Teichmann, C. et al., 2020: Assessing mean climate change signals in the global CORDEX-CORE ensemble. Climate Dynamics, doi:10.1007/s00382-020-05494-x.

Vautard, R., Kadygrov, N., Iles, C., Boberg, F., Buonomo, E., Bülow, K., et al. (2020). Evaluation of the large EURO-CORDEX regional climate model ensemble. Journal of Geophysical Research: Atmospheres, 125, e2019JD032344. Accepted Author Manuscript. https://doi.org/10.1029/2019JD032344

Widmann, M, Bedia, J, Gutiérrez, JM, et al. (2019) Validation of spatial variability in downscaling results from the VALUE perfect predictor experiment. Int J Climatol. 39: 3819–3845. https://doi.org/10.1002/joc.6024