GCM Selection & Ensemble Design: Best Practices and Recommendations from the EURO-CORDEX Community

Stefan Sobolowski[a,b], Samuel Somot[c], Jesus Fernandez[d], Guillaume Evin[e], Swen Brands[d], Douglas Maraun[f], Sven Kotlarski[g], Martin Jury[f], Rasmus E. Benestad[h], Claas Teichmann[i], Ole B. Christensen[j], Katharina Bülow[i], Erasmo Buonomo[k], Eleni Katragkou[l], Christian Steger[m], Silje Sørland[n], Grigory Nikulin[o], Carol McSweeney[k], Andreas Dobler[h], Tamzin Palmer[k], Renate Wilcke[o], Julien Boé[p], Lukas Brunner[q], Aurélien Ribes[c], Said Qasmi[c], Pierre Nabat[c], Florence Sevault[c], Thomas Oudar[r]

a. Geophysical Institute, University of Bergen, and the Bjerknes Center for Climate Research, Bergen Norway
b. NORCE Norwegian Research Centre, Bergen, Norway
c. Centre National de Recherches Météorologiques (CNRM), Université de Toulouse, Météo-France, CNRS, Toulouse, France
d. Instituto de Física de Cantabria (IFCA), CSIC-Universidad de Cantabria, Santander, Spain
e. University Grenoble Alpes, INRAE, CNRS, IRD, Grenoble INP, IGE, Grenoble, France
f. Wegener Center for Climate and Global Change, University of Graz, Graz, Austria.
g. Federal Office of Meteorology and Climatology (MeteoSwiss), Zurich, Switzerland
h. Norwegian Meteorological Institute, Oslo, Norway
i. Climate Service Center Germany (GERICS), Helmholtz-Zentrum Hereon, Hamburg, Germany
j. Danish Meteorological Institute (DMI), Copenhagen, Denmark
k. Met Office Hadley Centre, Exeter, UK
l. Aristotle University of Thessaloniki, School of Geology, Department of Meteorology and Climatology, Thessaloniki, Greece
m. Deutscher Wetterdienst, Offenbach, Germany
n. SWECO Norway AS, Bergen, Norway
o. Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden
p. CECI, Université de Toulouse, CERFACS, CNRS, Toulouse, France
q. Research Unit Sustainability and Climate Risk, Center for Earth System Research and Sustainability (CEN), Universität Hamburg, Hamburg, Germany
r. Météo-France, Toulouse, France

Corresponding author: Stefan Sobolowski, stefans@uib.no

Early Online Release: This preliminary version has been accepted for publication in *Bulletin of the American Meteorological Society*, may be fully cited, and has been assigned DOI 10.1175/BAMS-D-23-0189.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

High resolution climate information is critical for the vulnerability, impacts, adaptation, and climate services communities (VIACS). Coordinated ensembles generated by initiatives like CORDEX provide consistent and comparable information for the present and future over all land areas of the globe. This manuscript focuses on the European CORDEX initiative (hereafter EURO-CORDEX), and its coordinated effort to build regional climate ensembles for the years to come. In its first phase, EURO-CORDEX produced a rich ensemble of regional climate simulations under different representative concentration pathway scenarios. The EURO-CORDEX dataset is openly available and was fed into the Regional Atlas of 6th IPCC Assessment Report. However, this ensemble suffered from several shortcomings, which the community seeks to address in the next phase of production. Chief among these is the oft cited criticism that the selection of GCMs that provide input to the regional climate models was not rigorous and that the resulting ensemble represents an "ensemble of opportunity". The present paper provides a description of how the community has addressed these shortcomings. We present a comprehensive, flexible and traceable evaluation framework and toolkit for assessing the suitability of GCMs for downscaling, using EURO-CORDEX as an example. Its value lies in its explicit recognition of subjectivity and mechanisms implemented to transparently track decision making. Further, the utility of the framework extends well beyond pre-downscaling decisions to also include post-downscaling investigations performed by the VIACS communities and beyond, to include researchers investigating such topics as model biases, future constraints and exploring future storylines.

SIGNIFICANCE STATEMENT

The EURO-CORDEX community has created a comprehensive evaluation framework and open-source toolkit for global climate model assessment. These approaches provide a robust, transparent, traceable approach to both pre- and post-downscaling ensemble design whose utility extends well beyond the immediate regional climate community. This will enable improved and more nuanced assessments of regional climate change and its impacts.

CAPSULE (BAMS ONLY)

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

The EURO-CORDEX community presents an evaluation framework and open-source toolkit which can be employed to assess global climate model performance and build improved regional climate projections.

A. Introduction

Since the release of IPCC sixth assessment report there has been an increased awareness and focus on the local to regional scale responses to anthropogenically forced climate change (Masson-Delmotte et al., 2021). Indeed, within the report itself an entire chapter is dedicated to moving from global to regional scales (Doblas et al., 2021). Obtaining relevant information at these scales of interest typically requires downscaling (dynamical, statistical or hybrid). Here we focus on community downscaling efforts – using Regional Climate Models (RCMs) as well as Empirical Statistical Downscaling (ESD) – in the context of the WCRP-CORDEX initiative (Coordinated Regional Downscaling Experiments[1]). A longstanding challenge in downscaling (of all types) is the reliance on input from Global Climate Models (GCMs) that may have widely varying performance characteristics on regional scales, even though there has been steady improvement through CMIP generations (Brands, 2022a; Cannon, 2020). As such, efforts have sought to assess GCM performance to make well-informed choices prior to downscaling (McSweeney et al., 2012, 2015). Recent years have seen an increase in these efforts and an expansion of approaches with a particular focus on fitness for impacts assessments, adaptation and climate services applications (Ashfaq et al., 2022; Di Virgilio et al., 2022; Grose et al., 2023; Jeong & Cannon, 2023; Palmer et al., 2022). The different CORDEX domains have unique challenges of their own, due to their organization as a coordinated effort and their aim to build internally consistent and robust ensembles that capture as wide a range of future outcomes as possible. Here we focus on the European branch of CORDEX (hereafter, EURO-CORDEX) (Jacob et al., 2020) but aim to develop a general evaluation framework that is transparent, extensible and readily transferable to other domains and contexts. The framework and toolkit described here also provide an exciting opportunity for post hoc users of ensembles of downscaled data who may wish to perform additional tailoring of the ensemble to apply it to specific contexts.

1 https://cordex.org/

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

93 The main thrust of this manuscript and its accompanying toolkit, however, are meant to
94 assist a diverse community of modelers and users of regional climate data, both within
95 CORDEX and without, to make informed decisions when it comes to selecting GCMs for
96 downscaling from the CMIP6 archive. This work required finding consensus from a diverse
97 range of EURO-CORDEX community members. It is not always easy to agree to a single
98 approach and our work reflects the sometimes-competing interests community efforts must
99 confront. However, there is strong agreement on the overarching problem and its solution, as embodied by the following three points. First, while the approach to GCM selection during the CMIP5 downscaling phase cannot be wholly considered an "ensemble of opportunity" it is also true that the selection of GCMs was not as rigorous and transparent as it could have been. Second, we aim to improve upon this situation and help construct smarter, more reliable and more useful downscaled ensembles and make the selection process more objective and transparent. Third, the EURO-CORDEX community establishes this as a living approach that can evolve, and improve, along with our scientific understanding.

What follows are subsections describing the problem, the background and our ambitions. Following that, we outline the key categories and metrics that we deem important for evaluation of GCMs for downscaling and ultimately constructing an ensemble that covers the range of possible outcomes given the available scenarios (i.e. Shared Socioeconomic Pathways (SSPs)) (Sections B-F). We note that several decisions are taken that are context dependent; those surrounding SSPs and climate sensitivity reflect decisions that were important to the Euro-CORDEX community. Such decisions may reasonably vary for different applications and the general framework and toolkit presented here can be adapted depending on which choices one makes. Lastly, we discuss matrix[2] design considerations and statistical considerations (Sections G & H). We rely mostly on peer-reviewed literature for the assessment and evaluation of the CMIP6 simulations. These assessments are provided in the toolkit on GitHub, and we have devised innovative ways to include relevant metadata and update tables and spreadsheets with new results.[3]

Avoiding the "curse of opportunity"

2 The term "matrix" refers to the table of RCM-GCM pairings that form an ensemble. It often includes a third dimension for the scenarios.
3 https://wcrp-cordex.github.io/cmip6-for-cordex

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

While issues related to so-called "ensembles of opportunity" have been known for some time (Tebaldi and Knutti, 2007), there have been relatively few community-based efforts to address them. It arises when ensembles are constructed by asking for model results, in this case driving GCM data, from anyone who is willing to contribute. In the U.S., NARCCAP[4] (North American Regional Climate Change Assessment Program) used an experimental design, also referred to as 'factorial regression', to carefully design a matrix of a limited number of GCM-RCM combinations (Mearns et al., 2013). Building from NARCCAP, NA-CORDEX approached CMIP5 downscaling by considering factors such as Equilibrium Climate Sensitivity (ECS) and the quality of boundary conditions instead (see Bukovsky and Mearns 2020). The Southeast Asia CORDEX domain has developed a GCM ranking system that considers a handful of present-day performance metrics in CMIP6 models (Desmet & Ngo-Duc, 2022). It is not trivial, however, to pick a representative multi-model ensemble with different performance characteristics and indeed, no agreed upon, approach to determine which characteristics are "most" important (Benestad et al., 2017; Dalelane et al., 2018; Evin et al., 2021; Ferro et al., 2012; Knutti et al., 2009; Merrifield et al., 2023; Palmer et al., 2023). An additional challenge is how to ensure a subsample of a large multi-model ensemble of GCMs that maintains the model spread[5]. As a result, there are many different strategies for selecting ensembles, however, there is no "best" strategy.

These issues highlight that the design of ensembles and selection of driving models will always be fraught with trade-offs and subjective decisions. Rather than aim for the "perfect" or "best" models and ensemble designs, we aim to help researchers make well-informed choices and provide comprehensive information that will allow tailoring of ensembles to the purposes at hand (e.g. exploring storylines or worst-case scenarios). The contributors to this exercise acknowledge these tensions, the inherently subjective nature of some of these choices, and aim to make science-based decisions whenever possible. In addition, we wish to make the process transparent and comprehensible.

Background (EURO-CORDEX v1.0)

4 https://www.narccap.ucar.edu/
5 Provided there are no physical reasons for reducing it (e.g., observational constraints (Hegerl et al., 2021).

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

The currently available CMIP5-based EURO-CORDEX GCM-RCM-RCP matrix[6] was built up over 10 years and, generally, follows the CORDEX simulation protocol.[7] Acknowledging the general idea to sample both global/regional model uncertainty and emission scenario uncertainty (and not explicitly sampling internal climate variability[8]) simple approaches were applied, like using a few GCMs from all RCPs but many GCMs from one RCP, with special attention paid to sampling a wide range of GCM and RCP forcings.[9] Despite this, the CMIP5 EURO-CORDEX ensemble under sampled from the warmest and coldest GCMs and the driest GCMs (L. Coree, pers. comm.). Further, there was little rigorous assessment of the driving GCMs with respect to their performance (e.g., storm tracks, jet, SSTs & sea ice) and/or plausibility (e.g., ECS, trend reproduction) as was done *a posteriori* in e.g., McSweeney et al. (2012, 2015). The effort to select driving GCMs from the spread of future change, focused on changes in temperature and precipitation over Europe (Jury et al., 2015). These future change criteria are perhaps valid for temperature but not for precipitation, which is substantially modified by the RCM over the region of interest despite the large-scale dependencies. Lastly, there simply were not so many GCMs available in the CMIP5 archive that provided the requisite forcing data (Goldenson et al., 2023).

Consequently, the available EURO-CORDEX matrix provides an ensemble of impressive size and with a design that has at least been partly coordinated. However, in many respects the matrix still must be seen as an ensemble of opportunity with little assessment of driving GCMs' fitness and balance. This complicates, for instance, a complete use of the matrix in subsequent impact assessments. We therefore need to aim for a more consistent and coordinated design of the next (CMIP6-based) generation of the EURO-CORDEX matrix as well as a more robust assessment of the driving GCMs and their strengths and weaknesses, to progress as a community in "taking ensembles seriously and valuing model independence" (Jebeile & Barberousse, 2021).

Ambition for CMIP6 (EURO-CORDEX v2.0)

6 Tables of available Eur-11 (12km) simulations are available here: https://github.com/euro-cordex/esgf-table
7 While the discussion here is primarily concerned with the high resolution 12km ensemble, there is also an extensive ensemble of 50 km simulations. The CMIP5 protocol is here: https://cordex.org/wp-content/uploads/2020/05/cordex_general_instructions.pdf
8 This was true at least initially. However, more recently e.g., in PRINCIPLES (https://climate.copernicus.eu/c3s-production-european-climate-projections) internal variability has been included more explicitly. See also von Trentini et al. (2019).
9 https://climate.copernicus.eu/sites/default/files/2021-09/WebinarFAQ.pdf

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0189.1.

To address the above-mentioned issues, recent studies (published after the start of EURO-CORDEX) have explicitly proposed criteria to subsample the large CMIP5 GCM ensemble to serve a regional climate downscaling initiative (Brands et al., 2013; Jury et al., 2015; Parding et al., 2020) in addition to national climate service documents such as DRIAS, 2020[10] or the DWD "Referenz-/Kern Ensemble"[11]. Two overarching criteria have been emphasized by the authors, which we include in our framework:

1. Past-climate or baseline performance: Assuming that an RCM driven by a "well-performing" GCM will inherit the quality of the driver. Performance criteria have focused on fields relevant for RCM downscaling such as large-scale drivers (storm track position and intensity, weather regimes) and/or temperature seasonal cycle.
2. Future-climate spread: Assuming that an RCM driven by a highly sensitive GCM will be highly sensitive itself and vise-versa.

To our knowledge, the optimal way to subsample the shared socio-economic pathways or the individual members of a given GCM has not been specifically addressed in the literature though there is robust debate on the validity of some scenarios (Hausfather & Peters, 2020; Lawrence, 2020).

Following the lessons learnt from the literature and from previous RCM initiatives, we therefore propose the following four categories for GCM evaluation and selection for the CMIP6 EURO-CORDEX initiative:

- **Data availability:** GCM data is available to drive RCMs, and according to FAIR principles (Findable, Accessible, Interoperable, Reproducible).
- **GCM plausibility:** GCMs must be able realistically simulate key processes that drive climate in the region of interest, hereafter continental Europe.
- **Future climate change spread:** Once adequate performance is established, we aim to cover a range of future outcomes.
- **GCM independence:** Choose models in such a way that diversity is favored and avoid, as much as possible, redundancies.

10 In French, http://www.drias-climat.fr/accompagnement/sections/296
11 In German, https://www.dwd.de/DE/forschung/klima_umwelt/klimaprojektionen/fuer_deutschland/fuer_dtsl_rcp-datensatz_node.html

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

It is important to note that the aim here is not to exclude models (unless a major deficiency is noted) or create a ranking, but rather to inform the selection of GCMs for downscaling over Europe in as complete a manner as possible.

B. Data availability / quality

Our first category is data availability, basic quality control and FAIR principles. For this, we are thankful for the efforts of CORDEX-MIP (Gutowski Jr. et al., 2016), which gathered commitments from CMIP6 GCM teams to provide lateral boundary conditions necessary for dynamical and statistical downscaling. This is already an improvement over the *ad-hoc* approach to obtaining GCM output for downscaling in CMIP5. However, it should be noted that at the time of the initial data request the focus was on SSP1-2.6 and SSP5-8.5. This was before the CMIP6 simulations and their related SSPs had been analyzed in depth. It has since emerged that a more likely "business as usual" scenario is SSP3-7.0 (Hausfather & Peters, 2020). This was not widely known at the time of the CORDEX data request (2016-17) and as such this mismatch likely influenced data availability. Below are the criteria under this category.

Data availability criteria

1. CORDEX-MIP: This means that the necessary boundary conditions are provided
2. Basic Quality Assurance has been performed (QA e.g., missing values, suspect values, model levels, etc.)
3. The GCM data adheres to FAIR principles and is available on ESGF or similar (e.g., Climate Data Store)
4. GCMs provide data for a range of SSPs; in particular, SSP1-2.6, SSP2-4.5, SSP3-7.0 and SSP5-8.5 (with initial prioritization to be given to 2.6 and 7.0, to be consistent with the CMIP6 CORDEX experiment protocol)[12]
5. GCMs provide data necessary for both dynamical and statistical downscaling

Despite the efforts of CORDEX-MIP and the wide range of realizations and model versions (40 unique version/realizations combinations that meet sub-criteria 4 & 5) available, only 16 unique GCMs meet the above criteria. Given the independence criteria described below, the number of truly independent GCMs is likely smaller than this. The data

12 https://cordex.org/wp-content/uploads/2021/05/CORDEX-CMIP6_exp_design_RCM.pdf

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

availability for the different scenarios can be assessed by accessing the columns under the "1. Availability" heading in the toolkit.[13]

C. Plausibility criteria

We define "Plausibility criteria" as above under "GCM Plausibility". These criteria mainly focus on past and current climate performance but also include criteria on future climate responses such as TCR (we note this is a subjective choice and the framework is flexible in this regard). More generally, "Plausibility" should be understood in a broad sense in our general framework and other terms, such as "credibility" could also apply. Under this category, we establish performance-based criteria to select a group of global models that can reproduce key climate processes over the region of interest for downscaling. For a midlatitude area such as Europe this necessarily includes larger scale features of the climate system such as circulation over the North Atlantic, SSTs in the surrounding seas and sea ice in the Arctic. Because of this strong dependence on upstream conditions, we group our plausibility criteria under "Global/hemispheric" and "Regional" headings. The assumption is that realistic models will produce more realistic future projections, because they are able to represent processes correctly. For a detailed summary of the philosophy behind the application of such criteria see McSweeney et al. (2012) and Knutti et al. (2009). We also emphasize that the general framework and toolkit presented here are designed for flexibility and allow for cases where one may choose ***not*** to screen based on such metrics.

What follows is a list of the criteria used for selection of CMIP6 models. There are many others which might be included but the task team has endeavored to limit the criteria to processes or quantities that are of highest relevance for regional downscaling for the EURO-CORDEX domain (whether it be done via dynamical, statistical or hybrid approaches). Availability of published studies assessing many CMIP6 GCMs is also a limit for the plausibility criteria, knowing that new GCM assessment studies will likely emerge in the coming years. Where not available from published studies, the selected plausibility criteria have been calculated. We note that these are examples that are relevant for Europe and may be different for other regions. The toolkit allows for this flexibility.

Plausibility criteria

13 https://wcrp-cordex.github.io/cmip6-for-cordex/CMIP6_studies_table_EUR.html

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

Global

1) Observationally constrained transient climate response (TCR) (Ribes et al., 2021; Tokarska et al., 2020, IPCC-AR6)
2) Global performance scores (Brunner et al., 2020)
3) Observationally constrained global future climate change range for mid-21st century (Qasmi and Ribes, 2022)

Regional (Europe)[14]

1) Large-scale circulation criteria over the North-Atlantic such as Jet Stream North-South position (Oudar et al., 2020), storm track position (Pri20), blocking frequency (Dav20), Mean absolute error (MAE) for the frequency of European weather types (Brands 2022a) and a CMIP6 GCM revisited version of the McSweeney et al. (2015) criteria (Palmer et al., 2023)
2) Aerosol Optical Depth (AOD) Root Mean Square Error (RMSE) and past trend over Europe
3) Regional Sea Surface Temperature (SST) RMSE for surrounding water areas: Mediterranean (F. Sevault, pers. comm.), Black Sea, Norwegian and Barents Sea, Baltic and North Sea, North Atlantic Ocean
4) Regional Sea Ice Cover (SIC) RMSE for the relevant zones: Baltic Sea, Norwegian and Barents Sea

14 We note that a few key metrics are missing from the literature, these are: seasonal cycle of near-surface temperature, precipitation and humidity.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

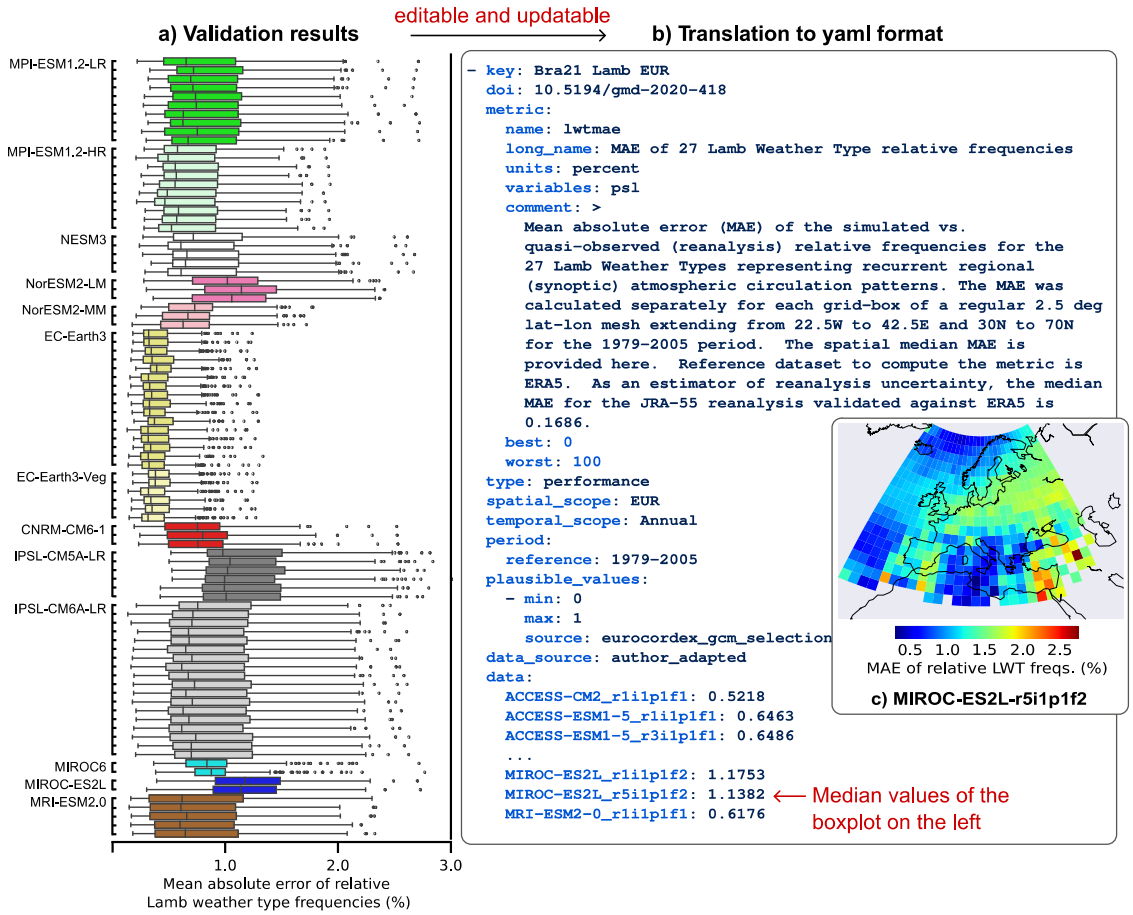

Figure 1. Summary GCM performance over the EURO-CORDEX domain for an example plausibility metric: the climatological (1979-2005) Lamb weather type frequencies as described in Brands (2022a). Error metric is the mean absolute error (MAE) of the 27 weather type frequencies simulated by a given GCM w.r.t. frequencies from ERA-5 reanalysis. MAE describes the GCM's capability to reproduce the near-surface atmospheric circulation centered at a particular grid-box; smaller MAE values indicate better performance, i.e. more plausible results. Each bar structure in panel a represents the error map (i.e. spatial sample) for an individual GCM run over the domain, the runs being indicated along the y-axis (the distinct runs of a given GCM are represented schematically). An example error map is shown in panel c for MIROC-ES2L-r5i1p1f2. Box in panel a = interquartile range (IQR) of the error sample; within-box black horizontal line = median value of the error sample - this value is listed in the Bra21.yaml file shown in panel b for each considered GCM run; boxplot whiskers in panel a are located at the first MAE value greater than p25 - 1.5 x IQR and at the last value less than p75 + 1.5 x IQR, with p25 and p75 = 25th and 75th percentile of the error sample, respectively. Similar color shadings refer to GCMs from the same institute. The yaml file shown in panel b is fully editable and updatable, see https://github.com/jesusff/cmip6-for-cordex/blob/main/CMIP6_studies/Bra21.yaml

The headings under the "2. Plausibility" section of the tables link to a file that includes a summary of the scores and plausibility limits, along with a link to the original publication (Figure 1). Taking Figure 1 as an example we can see that it is related to Lamb weather types. There are headings and comments explaining the scores, which in this example are the median MAE for a model and are displayed in the summary tables. Also shown in the file is the "plausibility range" which is set at 0-1. Models scoring over 1.0 are deemed implausible for this metric. The scores for other criteria vary but all appear in the summary tables, and it

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0189.1.

is clearly indicated whether they are within/without the acceptable range or above/below the chosen threshold. Note that many more studies might be available than those shown in the tables. Different criteria drove the decision to include/exclude specific studies. First, the main aim for the tables is to summarize the results, rather than being exhaustive. An exhaustive enumeration of all studies evaluating or presenting future outcomes of CMIP6 models would lead to unmanageably large tables. We promote a diversity of metrics exploring different aspects of model performance. Also, studies with a larger number of models and members are preferred, to avoid empty cells in the table (e.g., frequencies of lamb weather types). In any case, all studies considered are included in the GitHub repository and the causes for exclusion[15] from the summary table are recorded and the decision may be reversed after discussion or emergence of new evidence.

D. Selecting to cover the range of future outcomes

The third category in the framework for selection of driving global climate model simulations for downscaling is future projection coverage. Once the model simulations are available (Section B) and are shown to represent current and/or evolving climate reliably (Section C), we are left with a set of model simulations from which we can explore several plausible future climates. We recommend sampling these plausible futures using well defined uncertainty types as guidance:

- **Scenario uncertainty**

This type of uncertainty would be covered by selecting at least low/high concentration pathways. To obtain ensembles that are globally consistent, this choice has been agreed to for the whole CORDEX-CMIP6 initiative. SSP1-2.6 and SSP3-7.0 have been selected to illustrate these pathways. SSP2-4.5 has lower priority in the new CORDEX framework but could be used to simulate a medium pathway in agreement with current national efforts for reducing GHG emissions. Finally, the SSP5-8.5 is still of interest to illustrate extreme scenarios and worse-case trajectories and is especially of interest for those users conducting risk assessments. On a related note, this reasoning also applies to consideration of models with high TCR or ECS values (e.g., outside plausible ranges). Another option for handling

15 An example of a study (Fernandez-Granja et al., 2021) excluded from the table and how the causes for exclusion are registered can be found at https://github.com/WCRP-CORDEX/cmip6-for-cordex/blob/main/CMIP6_studies/Fer21.yaml

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

issues arising from scenario uncertainty and/or high sensitivity would be to apply a Global Warming Level approach, which is increasingly being used in impacts and adaptation contexts (Goldenson et al., 2023). The framework presented here can easily accommodate such modifications.

- **Model uncertainty**

For a given scenario, the model response to a prescribed external forcing is a key source of uncertainty, especially near the end of the 21st century (Evin et al., 2021; Hawkins & Sutton, 2009; Lehner et al., 2020). We need to carefully select the GCM variables or characteristics that are most relevant for the RCM simulations. Large-scale features such as weather regimes or North-Atlantic storm tracks are logical to explore. Exploring the model uncertainty in the GCM ensemble assumes that the RCM is not free to invent its own climate change signal within the domain and is constrained by the GCM sensitivity at least for some variables. This implies that RCMs are not disturbing the large-scale climate change signal, at least for some variables within the domain of interest. This is likely not true for all GCM-RCM pairs and the topic of GCM-RCM inconsistency is an active area of research (Boé et al., 2020; Taranu et al., 2022).

- **Internal variability**

Through the advent of Large Ensembles, the role of natural variability in modulating external forcing is now widely appreciated (Deser et al., 2012). And there have been some large ensemble efforts with regional climate models (Mote et al., 2016). It is important to take it into account, when possible, when designing GCM-RCM matrices and potentially when selecting specific member(s) for a given GCM. To our knowledge, dedicated work to the selection of GCM members has never been performed in the context of a multi-model coordinated effort such as CORDEX. This is partly related to the data availability issue as data required to drive RCMs are usually available only for one member per GCM except for some exceptions where internal and inter-model uncertainty are considered (von Trentini et al., 2019). In general, using more models has been considered more important than adding additional realizations of the same model (Longmate et al. 2023). This natural variability uncertainty is, however, intrinsically included into the RCM ensemble as random members of various GCMs are used. A few attempts to optimize the member selection have been

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

performed[16] but they went largely undocumented. This specific member selection is nevertheless relevant for the first decades of the scenario period to better cover the total climate change uncertainty range.

For CMIP6 there are 7 GCMs that make multiple realizations available for driving RCMs. However, we note that typically only one or two realizations have been evaluated for performance and other criteria in the literature. We believe it would be worthwhile, all other things being equal (i.e., the GCMs have satisfactory performance) to explore the role of internal variability explicitly by including these multiple realizations in the RCM-GCM matrix design of EURO-CORDEX. However, this has been challenging to achieve in practice.

Inspired by the literature and by the EURO-CORDEX RCM experiment protocol (Katragkou et al., 2024), we have considered the following criteria to explore the range of plausible futures. These also appear in the linked tables and are given a traffic light coloring to aid in interpretation and are displayed under the heading "3. Spread of future outcomes". We note that these are examples that are relevant for Europe and may be different for other regions. The toolkit allows for this flexibility.

Future change criteria

1) Jet stream position change (Oudar et al., 2020)
2) European near-surface temperature future change: 2070-2100 vs 1980-2010 for SSP5-8.5 (IPCC Atlas GitHub repo.) and observationally-constrained warming classes for JJA, 2041-2060 vs 1850-1900, SSP245 (Qasmi & Ribes, 2022)
3) Aerosol Optical Depth future evolution over Europe, SSP585, end of the 21st century (P. Nabat, pers. comm.)
4) Mediterranean SST future evolution, SSP585, end of the 21st century (F. Sevault, pers. comm.)
5) TCR values (IPCC-AR6) and ECS values (Schlund et al., 2020)

16 For instance, in EU-funded ENSEMBLES, ECHAM5-r3 was selected due to its better agreement with the observed trends. Also, the CNRM-CM5 member used in CORDEX-CMIP5 was initially the member r8i1p1 of the original ensemble and was renamed r1i1p1 by the GCM modelling group before diffusion on the ESGF and, in particular, for the provision of RCM LBCs. The specific choice of the member r8i1p1 was based on its better ability to reproduce 20th century past trends in global mean surface temperature (CNRM, pers. comm.).

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

As with the plausibility criteria the headings for the future change criteria in the tables include links to the underlying files.

E. Model independence and structural uncertainty

Here we introduce a final category of criteria, rarely discussed in the RCM-related literature to our knowledge but commonly used in practice: GCM independence (Boé, 2018; Brunner et al., 2020). Because GCMs are far from independent, the statistical properties (multi-model mean, standard deviation etc.) of the full ensemble may be quite biased if many interdependent models are considered in the driving GCM selection. We try here to assess the level of model dependency between the CMIP6 GCMs to eliminate obvious and less obvious near duplicates and to avoid introducing hidden biases in the ensemble and unnecessary duplication. For some applications the fine details will be important while for others simply knowing models share interdependencies is enough. We see two ways to assess and treat model independence:

- Independence criteria based on *a priori* model structure
- Independence criteria based on *a posteriori* model output pattern

Classifying models in families depending on their building phase, for example the relationship between institutes or the number of common lines of code or the list of common sub-models is a promising way to deal with independence. This a-priori model uncertainty or model independence criteria could rely on the model version (same model with different level complexity, concerning the spatial resolution, tuning choices or the number of climate components represented) or on the model lineage (different models with shared components i.e., shared lines of code). However, this *a priori* approach has been rarely attempted (Boé, 2018; Brands, 2022b; Leduc et al., 2016), probably because of the difficulties to obtain published, easy-to-handle and robust information about the models. The rising use of ES-DOC (https://search.es-doc.org/) may facilitate such an approach in the coming years and the GCM metadata archive built by Brands et al. (2023) provides useful information in the meantime.[17]

17 https://github.com/SwenBrands/gcm-metadata-for-cmip

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

Fortunately, model independence can also be assessed *a posteriori* by comparing the model outputs. In particular, the spatial pattern of the error maps and the future climate change response maps appear to align well with model dependency. This feature has been used first in Knutti, 2010 and Knutti & Sedláček, 2012 and more recently for the CMIP6 ensemble by Brunner et al. 2020 and Brands 2022b. Additionally we include an assessment of model complexity and the spatial resolution of the available models as the effective resolution of these models is typically much larger than their grid spacing. The criteria under this heading appear below and are also represented in the tables.

Model Independence / Structural uncertainty

1) Model complexity, by specifying the prescribed and interactive components considered (Brands 2022a)
2) Model Independence (Brunner et al., 2020; Brands 2022b)
3) Spatial resolution of a GCM e.g., effective resolution (Klaver et al., 2020)

While model independence was applied in the final evaluation and proposed selection (see Tables 1 & 2 below), we decided not to use the level of model complexity and the model resolution criteria as critical a-priori selection procedure, except as a tiebreaker, as they may duplicate other performance criteria. These criteria are shown in the summary tables of the toolkit under "4. Other criteria".

F. Approaches for merging criteria for reaching the final GCM list

Despite the aim to include all CMIP6 models in our assessment and acknowledging that even some "implausible" models may be useful to downscale for particular purposes, ultimately decisions on which models to prioritize for downscaling must be made. The accompanying tables are meant to facilitate this decision making. However, several questions persist that can only be answered by the community of practice itself and its specific context and not just by a subset of that community (i.e., the task team behind this manuscript). Below are some considerations:

- For which processes or scales is good performance most important for the community of practice?
- What are the thresholds for excluding models and how are they decided? (See subjectivity discussion above). 'Hard' and 'soft' thresholds can be more inclusive and

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

facilitate a nuanced construction of the ensemble more than an 'in' or 'out' criteria (see e.g., McSweeney et al 2012 & 2015, Palmer et al. 2023 and their use of a 'traffic light system).

- Do the different evaluation criteria carry equal weight? There may be clusters of non-independent criteria (e.g., jet stream latitude is related to blocking frequency).
- How should we best combine the plausibility criteria with the future spread criteria? E.g. McSweeney et al. 2012 use a clear 2-stage process whereby 1) models are evaluated and eliminated then 2) a diverse range of models is selected from those left. On the other hand, one could allow some poorly performing models 'in' because they cover a part of the future climate space that is underrepresented (See McSweeney et al., 2015 'decision making framework').

To facilitate the exploration of the decisions made, all background information and even the code used to create the tables are publicly available in the GitHub repository[18]. This includes human- and machine-readable files for every metric included in the tables and even those considered but not ultimately included (usually due to preferred, or more complete, sources for similar information). The source for the plausibility limits is also provided, since these are usually not stated in the published work and the authors were contacted to consider their input. This approach provides full transparency to the process of constructing the summary tables (see Figures 1-3). Moreover, it provides a way to update them in collaboration with the community, which can use GitHub issues to discuss their concerns with, provide alternative sources of information or even fork the whole repository to adapt it to their needs.

CMIP6 Model recommendations

Below we present tables where we summarize the details in the toolkit and provide a synthesis across all four evaluation categories and their underlying criteria (see Figure 1 for full details). These may be considered as an initial recommendation of which GCMs exhibit an appropriate level of performance for downscaling by the EURO-CORDEX community. However, this is a living process that can, and should, evolve in time (e.g. new evaluations are performed, additional simulations are added to ESGF).

18 https://github.com/WCRP-CORDEX/cmip6-for-cordex

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

All the models shown in Tables 1 & 2 satisfy the availability criteria. The plausibility scores are summarized in the "Marks/Criteria" column. A model gets a mark for a given criterion if it lies outside the defined range or above/below a defined threshold. Again, all these decisions are detailed in the accompanying yaml files (see Figures 2 & 3 for examples). The reader is encouraged to visit the toolkit website for the details as it is not practical to display all the information here. The strictest of these tables (Table 1) is an illustration of the fact that being too restrictive on the criteria likely leads to too few models to ensure reliable downscaled ensembles. It should be noted that we applied a threshold so that models which score highly simply because they have not been evaluated are not included (a minimum of one plausibility score is required). We also note that being evaluated for multiple criteria should be viewed favorably and that all the displayed models were evaluated for at least 15 plausibility criteria (e.g. Global warming level, large-scale circulation, surface forcings such SST, SIC or AOD). For the future change category, we chose the TCR criterion to organize the models according to warming levels. There are many other possible configurations that can be constructed using the filtering features in the GitHub tables. Here we present two, the second of which we believe represents an appropriate starting point for CMIP6 downscaling in EURO-CORDEX.[19]

Table 1. Strictest; GCMs which are available for all four scenarios (ssp126, ssp245, ssp370,ssp585) and are deemed "plausible" for all evaluated criteria. To qualify models must be evaluated for at least one criterion per category (availability, plausibility, future change, independence). The third column shows the number of failed criteria over the total number of criteria for each model. Models that are also part of institutional commitments are highlighted. The fourth column shows an illustration of future spread categories for the selected GCMs (here based on TCR values from low(green) to high (red)).

GCM name	Run	Marks/Criteria	TCR Plausible range (1.2K-2.4K)[20]
MPI-ESM1-2-LR	r1i1p1f1	0/18	1.84

Table 2. Less strict; same as Table 1 except for removing the "plausible for all evaluated criteria" argument / filter. Scores are based on all evaluated members of a model even if only one member is "available". Only one model per family is kept in most cases and in the event of a tie criteria such as complexity and resolution may play a role as tiebreakers. Explanations appear in footnotes.

19 https://wcrp-cordex.github.io/cmip6-for-cordex/CMIP6_studies_table_EUR.html.
20 TCR 90% range as provided by the IPCC AR6
https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Chapter_07_Supplementary_Material.pdf

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

GCM name	Run	Marks/Criteria	TCR Plausible range (1.2K-2.4K)
NorESM2-MM[21]	r1i1p1f1	1/17	1.33
MIROC6[22]	r1i1p1f1	1/20	1.55
MPI-ESM1-2-HR	r1i1p1f1	1/20	1.66
CNRM-ESM2-1	r1i1p1f2	1/19	1.86
CESM2[23]	r11i1p1f1	1/18	2.06
CMCC-CM2-SR5[24]	r1i1p1f1	1/15	2.09
IPSL-CM6A-LR[25]	r1i1p1f1	2/16	2.32
EC-Earth3-Veg[26]	r1i1p1f1	2/15	2.62
UKESM1-0-LL[27]	r1i1p1f2	2/19	2.79

Two additional models that are part of institutional commitments are *not* currently "available". These are: EC-EARTH3-Veg r6i1p1f1 and EC-EARTH3 r1i1p1f1. It is likely these will appear on the ESGF well before EURO-CORDEX simulations are completed. As noted above this list will evolve as additional analyses are conducted. Models that currently score well in terms of performance may be downgraded as more analyses are performed. This is especially a concern for those which have only been evaluated for a few criteria. As such we focus on those evaluated for many criteria. Likewise, models/simulations that are

21 Same family as CESM but helps with covering future spread as it has low TCR. NorESM2-MM performs much better than NorESM2-LM in terms of regional atmospheric circulation (Brands 2022a)
22 MIROC-ES2L might also be considered, but has a 500km nominal resolution, scores ⅗ and is the same family as MIROC6.
23 TaiESM1 would also qualify but shares all components with CESM2 (https://gmd.copernicus.org/articles/13/3887/2020/).
24 CMCC-ESM2 might also be considered but it is evaluated for fewer criteria, is the same model family and is not evaluated for TCR. CMCC-CM2-SR5 shares its AGCM (CAM5.3), LSM (CLM4.5) and sea-ice model (CICE4.0) with CESM1.
25 ISPL-CM6A-LR shares its OGCM (NEMO3.6) and ocean biogeochemistry model (PISCESv2) with CNRM-ESM2-1. However, it is likely that there will be an institutional commitment to downscale the IPSL model (t.b.c).
26 EC-EARTH3 would also qualify but shares the same family.
27 ACCESS-CM2 also qualifies but is from the same family and does not include the Black Sea which could present issues for the EURO-CORDEX domain.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

currently unavailable on ESGF may become available. Further, marks against plausibility should not be automatically disqualifying. Rather, they should be considered as flags that warn of known shortcomings that downscaling teams should consider carefully. As an example, we point to the ACCESS models, which perform well but one of their marks against, arises due to the absence of the Black Sea. This omission may have serious implications for the southeast Mediterranean and eastern Europe. While it might be tempting to turn the Marks/Criteria ratios into a skill-score many of the criteria cannot be considered independent and such a skill-score would need to be carefully constructed to avoid spurious results. Nevertheless, we can see that the models in Table 2. generally perform quite well. They also cover well the range of plausible TCR[28] with a few warm outliers, which may be of interest for climate service providers aiming to provide risk assessments based on worst-case scenarios (EC-EARTH and UKESM).

G. Matrix design

A final issue to consider for the robust construction of downscaled ensembles of CMIP6, is how best to fill the 3-D matrix of simulations (either RCM-GCM-SSP or ESD-GCM-SSP pairings). As noted in the introduction, the matrix of simulations from CMIP5 remains sparse, heterogenous and unbalanced despite some impressive efforts to ameliorate these shortcomings. It is very large and presents challenges for impact modelers and climate service providers who may wish to perform a sub-selection of the ensemble, often without sufficient guidance. This challenge will persist in the downscaling of CMIP6 due to differences in institutional strategies, resources, timing etc. In some ways the EURO-CORDEX ensembles are, like CMIP ensembles, often unbalanced as they grow over time and are "dynamic" in practice, if not by design. However, we believe we can take some steps at the outset that can help reduce the impact of typical matrix design hazards and agree to a "balanced matrix" experiment (that exists as a subset of the much larger EURO-CORDEX ensemble) that can be used in adaptation planning and impacts assessments.

We suggest four criteria to help guide filling the matrix:

1) Balance the ensemble as best as possible in order that no SSP, GCM or RCM are over- or under-represented.

28 TCR is just one possibility, other spread criteria could also be used here.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

2) Physical consistency between GCM and RCM, so that the RCM does not deviate too much from the GCMs at large-scales.
3) Keeping an eye on the range of uncertainty. For example, we should not artificially decrease the range of outcomes provided by the GCMs and should allow the possibility to explore interesting but rare outcomes such as plausible worst-case scenarios.
4) Facilitate a-posteriori filling of the matrix via statistical approaches (see next subsection).

Table 3 shows the selection of ongoing or planned EURO-CORDEX RCM simulations to date that contribute to a balanced matrix that fulfills the criteria mentioned above. Efforts have been made to avoid under/overrepresented GCMs and represent, as much as possible, a fractional factorial design where at least 3 runs by RCM and 4 runs by GCM are considered. However, there are also institutional or project constraints that must be respected, which leads to some GCMs that are more represented than other ones. We note that all the GCMs satisfy the evaluation criteria detailed in the previous sections.

Table 3. The EURO-CORDEX balanced matrix experiment comprises simulations filling a GCM-RCM-SSP combination matrix with a fractional factorial design where at least 3 runs by RCM and 4 runs by GCM are considered (two additional simulations with CMCC-CM2-SR5 and MIROC6 need to be planned to fit this criteria). Blue crosses indicate the simulations which are planned, green crosses indicate simulations which are currently (Feb. 2023) running or are completed[29]. All simulations are planned to run with the scenarios SSP1-2.6 and SSP3-7.0.

	Institution / RCM						
Driving GCM / run	**CNRM / ALADIN6x**	**CLMcom / ICON-CLM**	**HCLIMcom / HCLIM43 - ALADIN**	**KNMI/ RACMO23E**	**GERICS / REMO**	**CUNI, ICTP / RegCM**	**AUTH, CESAM,.. / WRF**
NorESM2-MM / r1i1p1f1	✘			✘		✘	✘
MIROC6 / r1i1p1f1		✘	✘		✘		

29 Updated status information can be found in https://wcrp-cordex.github.io/simulation-status/CORDEX_CMIP6_status_by_experiment.html#EUR-11-EURbalanced

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

MPI-ESM1-2-HR / r1i1p1f1		✘			✘	✘	✘
CNRM-ESM2-1 / r1i1p1f2	✘		✘	✘		✘	
CMCC-CM2-SR5 / r1i1p1f1	✘	✘					✘
EC-Earth3-Veg / r1i1p1f1			✘	✘	✘	✘	

H. Statistical exploitation of the EURO-CORDEX ensemble

There are several ways to make full use of the EURO-CORDEX ensemble and we discuss a few here. For example, combining dynamically downscaled results with empirical-statistical downscaling (ESD) which can be applied to a much larger multi-model ensemble, and can take advantage of relative strengths while mitigating weaknesses (Mezghani et al., 2019). We have noted that we try to avoid an ensemble of opportunity in the EURO-CORDEX downscaled ensemble. However, the input GCM ensemble (CMIP6) is itself an ensemble of opportunity in the sense that there are some model lineages/versions which are over- represented in the ensemble. CMIP6 may also be under dispersive of the full range of outcomes due to scenario considerations, model uncertainty and/or internal variability. Even though ESD approaches can downscale the full CMIP ensemble, a weighting scheme or thinning of the full ensemble would be required to have a balanced ensemble of downscaled results. In an unbalanced multi-model ensemble, not weighting is weighting (Fernández & Frías, 2020). Statistical approaches could be used to assess the effects of different ensemble thinning strategies *a posteriori* (Christensen & Kjellström, 2020) to statistically obtain balanced climate change signal and the associated uncertainties with an a-posteriori statistical filling of the matrix (Déqué et al., 2012; Evin et al., 2021). A complementary approach involves so-called “hybrid” downscaling techniques (Doblas-Reyes et al., 2021). For aggregated variables, such as the mean temperature or rainfall totals, it may be possible to use hybrid downscaling to emulate GCM-RCM pairs and hence extend the results of a selected RCM to a large ensemble of GCMs (Erlandsen et al., 2020). This involves using existing GCM-RCM pairs for calibrating statistical models which are subsequently applied to

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

different GCMs. Statistical emulation is now even possible at the daily scale, allowing one to potentially mimic a time series of 2D maps of the missing GCM-RCM pairs (Doury et al., 2023). Emulation opens the door to adoption of machine learning methods and potentially pushing to higher spatial resolutions in dynamical downscaling (e.g., convection permitting scales) as fewer simulations would be needed. This is a nascent but rapidly developing area of research that nevertheless holds great promise.

I. Conclusions

This manuscript, its accompanying meta-analyses and recommendations (see Table 2) reflects an effort by the EURO-CORDEX community to provide a more robust and transparent basis for the selection of driving GCMs and approaches to filling the matrix of simulations. There is no attempt here to point to a "best" performing set of models. Rather, we seek to provide guidance via a comprehensive assessment, backed by the available literature, of the CMIP6 GCMs under consideration for dynamical downscaling. As discussed above, we are aware that any selection of GCMs is inherently subjective, but we aim to ensure that the criteria applied are supported by results from as many experts from the regional and global climate modelling communities as possible i.e., to make the final decision as "objective" as possible. Furthermore, the criteria and the decision process should be documented to guarantee full transparency, traceability and verifiability. We believe that a science and information driven selection/guidance approach is in any case an improvement compared to the uncoordinated and, to a certain degree, coincidental selection procedure that has been applied in the past.

This manuscript articulates a general framework for GCM evaluation and a general toolkit to facilitate this. It also provides a generalizable approach to ensemble design (see section G) that can address some of the pitfalls involved in a posteriori ensemble member subselection. In these respects the manuscript differs from the white paper that formed its basis and is more of an implementation document specific to the Euro-CORDEX community (Sobolowski et al., 2023). The task team welcomes suggestions for improvements and additions to the tools and documentation. In particular, we emphasize that the choices made from the EURO-CORDEX perspective may not be the most relevant for other regions and/or use cases. The toolkit is flexible in this regard. It has been designed in a way that allows its straightforward application to other CORDEX domains (it is currently extended to the

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

Southeast Asia, Australasia and Mediterranean domains), other downscaling initiatives and even investigations such as physical climate storylines.

Acknowledgments.

SS acknowledges support from the education ministry of Norway via its base funding of the Bjerknes Center for Climate Research as well as Horizon Europe project, Impetus4Change (grant nr. 101081555); JF acknowledges support from project CORDyS (PID2020-116595RB-I00) funded by MCIN/AEI/10.13039/501100011033; EB was supported by the UK's FCDO funded Weather and Climate Information Services (WISER) programme; MJ acknowledges funding from the Austrian Climate and Energy Fund under its 14th call of the Austrian Climate Research Programme, project PREVAL ÖKS NEXTGEN (project number C265151); SB acknowledges funding by CSIC's Interdisciplinary Thematic Platform Clima (PTI-Clima), supported by the Spanish Ministry for Ecological Transition and Demographic Challenge (MITECO) and the NextGenerationEU recovery plan of the European Commission (Regulation EU 2020/2094). EK acknowledges funding from the Hellenic Foundation for Research & Innovation (HFRI), under Sub-action 2-Funding Projects in Leading-Edge Sectors – RRFQ: Basic Research Financing (Horizontal support for all Sciences), Proj Nr: 014696. S. Somot has received funding from Agence Nationale de la Recherche - France 2030 as part of the PEPR TRACCS programme under grant number ANR-22-EXTR-00XX (LOCALISING).

Data Availability Statement.

All data and analyses used in this study are available here: https://wcrp-cordex.github.io/cmip6-for-cordex

References

Ashfaq, M., Rastogi, D., Kitson, J., Abid, M. A., & Kao, S.-C. (2022). Evaluation of CMIP6 GCMs Over the CONUS for Downscaling Studies. *Journal of Geophysical Research: Atmospheres*, *127*(21), e2022JD036659. https://doi.org/10.1029/2022JD036659

Benestad, R., Sillmann, J., Thorarinsdottir, T. L., Guttorp, P., Mesquita, M. d S., Tye, M. R., Uotila, P., Maule, C. F., Thejll, P., Drews, M., & Parding, K. M. (2017). New vigour involving statisticians to overcome ensemble fatigue. *Nature Climate Change*, *7*(10), 697–703. https://doi.org/10.1038/nclimate3393

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

Boé, J. (2018). Interdependency in Multimodel Climate Projections: Component Replication and Result Similarity. *Geophysical Research Letters*, *45*(6), 2771–2779. https://doi.org/10.1002/2017GL076829

Boé, J., Somot, S., Corre, L., & Nabat, P. (2020). Large discrepancies in summer climate change over Europe as projected by global and regional climate models: Causes and consequences. *Climate Dynamics*, *54*(5–6), 2981–3002. https://doi.org/10.1007/s00382-020-05153-1

Brands, S. (2022a). A circulation-based performance atlas of the CMIP5 and 6 models for regional climate studies in the Northern Hemisphere mid-to-high latitudes. *Geoscientific Model Development*, *15*(4), 1375–1411. https://doi.org/10.5194/gmd-15-1375-2022

Brands, S. (2022b). Common Error Patterns in the Regional Atmospheric Circulation Simulated by the CMIP Multi-Model Ensemble. *Geophysical Research Letters*, *49*(23). https://doi.org/10.1029/2022GL101446

Brands, S., Herrera, S., Fernández, J., & Gutiérrez, J. M. (2013). How well do CMIP5 Earth System Models simulate present climate conditions in Europe and Africa?: A performance comparison for the downscaling community. *Climate Dynamics*, *41*(3–4), 803–817. https://doi.org/10.1007/s00382-013-1742-8

Brands, S., Tatebe, H., Danek, C., Fernández, J., Swart, N. C., Volodin, E., Kim, Y., Collier, M., Bi, D., & Tongwen, W. (2023). A metadata archive about global climate model complexity in CMIP (v1.1). Zenodo. https://doi.org/10.5281/zenodo.7813495

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., & Knutti, R. (2020). Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth System Dynamics*, *11*(4), 995–1012. https://doi.org/10.5194/esd-11-995-2020

Bukovsky, M. S., & Mearns, L. O. (2020). Regional climate change projections from NA-CORDEX and their relation to climate sensitivity. *Climatic Change*, *162*(2), 645–665. https://doi.org/10.1007/s10584-020-02835-x

Cannon, A. J. (2020). Reductions in daily continental-scale atmospheric circulation biases between generations of global climate models: CMIP5 to CMIP6. *Environmental Research Letters*, *15*(6), 064006. https://doi.org/10.1088/1748-9326/ab7e4f

Christensen, O. B., & Kjellström, E. (2020). Partitioning uncertainty components of mean climate and climate change in a large ensemble of European regional climate model projections. *Climate Dynamics*, *54*(9), 4293–4308. https://doi.org/10.1007/s00382-020-05229-y

Dalelane, C., Früh, B., Steger, C., & Walter, A. (2018). A Pragmatic Approach to Build a Reduced Regional Climate Projection Ensemble for Germany Using the EURO-CORDEX 8.5 Ensemble. *Journal of Applied Meteorology and Climatology*, *57*(3), 477–491. https://doi.org/10.1175/JAMC-D-17-0141.1

Déqué, M., Somot, S., Sanchez-Gomez, E., Goodess, C. M., Jacob, D., Lenderink, G., & Christensen, O. B. (2012). The spread amongst ENSEMBLES regional scenarios: Regional climate models, driving general circulation models and interannual variability. *Climate Dynamics*, *38*(5), 951–964. https://doi.org/10.1007/s00382-011-1053-x

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

Deser, C., Knutti, R., Solomon, S., & Phillips, A. S. (2012). Communication of the role of natural variability in future North American climate. *Nature Climate Change*, *2*(11), nclimate1562. https://doi.org/10.1038/nclimate1562

Desmet, Q., & Ngo-Duc, T. (2022). A novel method for ranking CMIP6 global climate models over the southeast Asian region. *International Journal of Climatology*, *42*(1), 97–117. https://doi.org/10.1002/joc.7234

Di Virgilio, G., Ji, F., Tam, E., Nishant, N., Evans, J. P., Thomas, C., Riley, M. L., Beyer, K., Grose, M. R., Narsey, S., & Delage, F. (2022). Selecting CMIP6 GCMs for CORDEX Dynamical Downscaling: Model Performance, Independence, and Climate Change Signals. *Earth's Future*, *10*(4), e2021EF002625. https://doi.org/10.1029/2021EF002625

Doblas-Reyes, F., Sörensson, A., Almazroui, M., Dosio, A., Gutowski, W., Haarsma, R., Hamdi, R., Hewitson, B., Kwon, W.-T., Lamptey, B., Maraun, D., Stephenson, T., Takayabu, I., Terray, L., Turner, A., & Zuo, Z. (2021). *IPCC AR6 WGI Chapter 10: Linking global to regional climate change* (pp. 1363–1512). https://doi.org/10.1017/9781009157896.012

Doury, A., Somot, S., Gadat, S., Ribes, A., & Corre, L. (2023). Regional climate model emulator based on deep learning: Concept and first evaluation of a novel hybrid downscaling approach. *Climate Dynamics*, *60*(5), 1751–1779. https://doi.org/10.1007/s00382-022-06343-9

Erlandsen, H. B., Parding, K. M., Benestad, R., Mezghani, A., & Pontoppidan, M. (2020). A Hybrid Downscaling Approach for Future Temperature and Precipitation Change. *Journal of Applied Meteorology and Climatology*, *59*(11), 1793–1807. https://doi.org/10.1175/JAMC-D-20-0013.1

Evin, G., Somot, S., & Hingray, B. (2021). Balanced estimate and uncertainty assessment of European climate change using the large EURO-CORDEX regional climate model ensemble. *Earth System Dynamics Discussions*, 1–40. https://doi.org/10.5194/esd-2021-8

Fernández, J., & Frías, M. D. (2020). *Balanced subsampling of future regional climate ensembles of opportunity* [Other]. oral. https://doi.org/10.5194/egusphere-egu2020-20052

Ferro, C. A. T., Jupp, T. E., Lambert, F. H., Huntingford, C., & Cox, P. M. (2012). Model complexity versus ensemble size: Allocating resources for climate prediction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *370*(1962), 1087–1099. https://doi.org/10.1098/rsta.2011.0307

Goldenson, N., Leung, L. R., Mearns, L. O., Pierce, D. W., Reed, K. A., Simpson, I. R., Ullrich, P., Krantz, W., Hall, A., Jones, A., & Rahimi, S. (2023). Use-Inspired, Process-Oriented GCM Selection: Prioritizing Models for Regional Dynamical Downscaling. *Bulletin of the American Meteorological Society*, *1*(aop). https://doi.org/10.1175/BAMS-D-23-0100.1

Grose, M. R., Narsey, S., Trancoso, R., Mackallah, C., Delage, F., Dowdy, A., Di Virgilio, G., Watterson, I., Dobrohotoff, P., Rashid, H. A., Rauniyar, S., Henley, B., Thatcher, M., Syktus, J., Abramowitz, G., Evans, J. P., Su, C.-H., & Takbash, A. (2023). A CMIP6-based multi-model downscaling ensemble to underpin climate change

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

services in Australia. *Climate Services*, *30*, 100368. https://doi.org/10.1016/j.cliser.2023.100368

Gutowski Jr., W. J., Giorgi, F., Timbal, B., Frigon, A., Jacob, D., Kang, H.-S., Raghavan, K., Lee, B., Lennard, C., Nikulin, G., O'Rourke, E., Rixen, M., Solman, S., Stephenson, T., & Tangang, F. (2016). WCRP COordinated Regional Downscaling EXperiment (CORDEX): A diagnostic MIP for CMIP6. *Geosci. Model Dev.*, *9*(11), 4087–4095. https://doi.org/10.5194/gmd-9-4087-2016

Hausfather, Z., & Peters, G. P. (2020). Emissions – the 'business as usual' story is misleading. *Nature*, *577*(7792), 618–620. https://doi.org/10.1038/d41586-020-00177-3

Hawkins, E., & Sutton, R. (2009). The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1107. https://doi.org/10.1175/2009BAMS2607.1

Hegerl, G. C., Ballinger, A. P., Booth, B. B. B., Borchert, L. F., Brunner, L., Donat, M. G., Doblas-Reyes, F. J., Harris, G. R., Lowe, J., Mahmood, R., Mignot, J., Murphy, J. M., Swingedouw, D., & Weisheimer, A. (2021). Toward Consistent Observational Constraints in Climate Predictions and Projections. *Frontiers in Climate*, *3*. https://www.frontiersin.org/articles/10.3389/fclim.2021.678109

Jacob, D., Teichmann, C., Sobolowski, S., Katragkou, E., Anders, I., Belda, M., Benestad, R., Boberg, F., Buonomo, E., Cardoso, R. M., Casanueva, A., Christensen, O. B., Christensen, J. H., Coppola, E., De Cruz, L., Davin, E. L., Dobler, A., Domínguez, M., Fealy, R., … Wulfmeyer, V. (2020). Regional climate downscaling over Europe: Perspectives from the EURO-CORDEX community. *Regional Environmental Change*, *20*(2), 51. https://doi.org/10.1007/s10113-020-01606-9

Jebeile, J., & Barberousse, A. (2021). Model spread and progress in climate modelling. *European Journal for Philosophy of Science*, *11*(3), 66. https://doi.org/10.1007/s13194-021-00387-0

Jeong, D. I., & Cannon, A. J. (2023). An Approach for Selecting Observationally-Constrained Global Climate Model Ensembles for Regional Climate Impacts and Adaptation Studies in Canada. *Atmosphere-Ocean*, *0*(0), 1–17. https://doi.org/10.1080/07055900.2023.2239194

Jury, M. W., Prein, A. F., Truhetz, H., & Gobiet, A. (2015). Evaluation of CMIP5 Models in the Context of Dynamical Downscaling over Europe. *Journal of Climate*, *28*(14), 5575–5582. https://doi.org/10.1175/JCLI-D-14-00430.1

Katragkou, E., Sobolowski, S. P., Teichmann, C., Solmon, F., Pavlidis, V., Rechid, D., Hoffmann, P., Fernandez, J., Nikulin, G., & Jacob, D. (2024). Delivering an Improved Framework for the New Generation of CMIP6-Driven EURO-CORDEX Regional Climate Simulations. *Bulletin of the American Meteorological Society*, *105*(6), E962–E974. https://doi.org/10.1175/BAMS-D-23-0131.1

Klaver, R., Haarsma, R., Vidale, P. L., & Hazeleger, W. (2020). Effective resolution in high resolution global atmospheric models for climate studies. *Atmospheric Science Letters*, *21*(4), e952. https://doi.org/10.1002/asl.952

Knutti, R. (2010). The end of model democracy? *Climatic Change*, *102*(3), 395–404. https://doi.org/10.1007/s10584-010-9800-2

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2009). Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate*, *23*(10), 2739–2758. https://doi.org/10.1175/2009JCLI3361.1

Knutti, R., & Sedláček, J. (2012). Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change*, *3*(4), nclimate1716. https://doi.org/10.1038/nclimate1716

Lawrence, D. (2020). Uncertainty introduced by flood frequency analysis in projections for changes in flood magnitudes under a future climate in Norway. *Journal of Hydrology: Regional Studies*, *28*, 100675. https://doi.org/10.1016/j.ejrh.2020.100675

Leduc, M., Laprise, R., Elía, R. de, & Šeparović, L. (2016). Is Institutional Democracy a Good Proxy for Model Independence? *Journal of Climate*, *29*(23), 8301–8316. https://doi.org/10.1175/JCLI-D-15-0761.1

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E. M., Brunner, L., Knutti, R., & Hawkins, E. (2020). Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth System Dynamics*, *11*(2), 491–508. https://doi.org/10.5194/esd-11-491-2020

Longmate, J.M., Risser, M.D. & Feldman, D.R. Prioritizing the selection of CMIP6 model ensemble members for downscaling projections of CONUS temperature and precipitation. Clim Dyn 61, 5171-5197 (2023). https://doi.org/10.1007/s00382-023-06846-z

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, Ö., Yu, R., & Zhou, B. (Eds.). (2021). Annex X: Expert Reviewers of the IPCC Working Group I Sixth Assessment Report. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 2287–2338). Cambridge University Press. https://doi.org/10.1017/9781009157896.001

McSweeney, C. F., Jones, R. G., & Booth, B. B. B. (2012). Selecting Ensemble Members to Provide Regional Climate Change Information. *Journal of Climate*, *25*(20), 7100–7121. https://doi.org/10.1175/JCLI-D-11-00526.1

McSweeney, C. F., Jones, R. G., Lee, R. W., & Rowell, D. P. (2015). Selecting CMIP5 GCMs for downscaling over multiple regions. *Climate Dynamics*, *44*(11), 3237–3260. https://doi.org/10.1007/s00382-014-2418-8

Mearns, L. O., Sain, S., Leung, L. R., Bukovsky, M. S., McGinnis, S., Biner, S., Caya, D., Arritt, R. W., Gutowski, W., Takle, E., Snyder, M., Jones, R. G., Nunes, A. M. B., Tucker, S., Herzmann, D., McDaniel, L., & Sloan, L. (2013). Climate change projections of the North American Regional Climate Change Assessment Program (NARCCAP). *Climatic Change*, *120*(4), 965–975. https://doi.org/10.1007/s10584-013-0831-3

Merrifield, A. L., Brunner, L., Lorenz, R., Humphrey, V., & Knutti, R. (2023). Climate model Selection by Independence, Performance, and Spread (ClimSIPS) for regional applications. *EGUsphere*, 1–49. https://doi.org/10.5194/egusphere-2022-1520

Mezghani, A., Dobler, A., Benestad, R., Haugen, J. E., Parding, K. M., Piniewski, M., & Kundzewicz, Z. W. (2019). Subsampling Impact on the Climate Change Signal over

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

Poland Based on Simulations from Statistical and Dynamical Downscaling. *Journal of Applied Meteorology and Climatology*, *58*(5), 1061–1078. https://doi.org/10.1175/JAMC-D-18-0179.1

Mote, P. W., Allen, M. R., Jones, R. G., Li, S., Mera, R., Rupp, D. E., Salahuddin, A., & Vickers, D. (2016). *Superensemble Regional Climate Modeling for the Western United States*. https://doi.org/10.1175/BAMS-D-14-00090.1

Oudar, T., Cattiaux, J., & Douville, H. (2020). Drivers of the Northern Extratropical Eddy-Driven Jet Change in CMIP5 and CMIP6 Models. *Geophysical Research Letters*, *47*(8), e2019GL086695. https://doi.org/10.1029/2019GL086695

Palmer, T. E., McSweeney, C. F., Booth, B. B. B., Priestley, M. D. K., Davini, P., Brunner, L., Borchert, L., & Menary, M. B. (2022). Performance based sub-selection of CMIP6 models for impact assessments in Europe. *Earth System Dynamics Discussions*, 1–45. https://doi.org/10.5194/esd-2022-31

Palmer, T. E., McSweeney, C. F., Booth, B. B. B., Priestley, M. D. K., Davini, P., Brunner, L., Borchert, L., & Menary, M. B. (2023). Performance-based sub-selection of CMIP6 models for impact assessments in Europe. *Earth System Dynamics*, *14*(2), 457–483. https://doi.org/10.5194/esd-14-457-2023

Parding, K. M., Dobler, A., McSweeney, C. F., Landgren, O. A., Benestad, R., Erlandsen, H. B., Mezghani, A., Gregow, H., Räty, O., Viktor, E., El Zohbi, J., Christensen, O. B., & Loukos, H. (2020). GCMeval – An interactive tool for evaluation and selection of climate model ensembles. *Climate Services*, *18*, 100167. https://doi.org/10.1016/j.cliser.2020.100167

Qasmi, S., & Ribes, A. (2022). Reducing uncertainty in local temperature projections. *Science Advances*, *8*(41), eabo6872. https://doi.org/10.1126/sciadv.abo6872

Ribes, A., Qasmi, S., & Gillett, N. P. (2021). Making climate projections conditional on historical observations. *Science Advances*, *7*(4), eabc0671. https://doi.org/10.1126/sciadv.abc0671

Schlund, M., Lauer, A., Gentine, P., Sherwood, S. C., & Eyring, V. (2020). Emergent constraints on equilibrium climate sensitivity in CMIP5: Do they hold for CMIP6? *Earth System Dynamics*, *11*(4), 1233–1258. https://doi.org/10.5194/esd-11-1233-2020

Sobolowski, Stefan, Somot, Samuel, Fernandez, Jesus, Evin, Guillaume, Maraun, Douglas, Kotlarski, Sven, Jury, Martin, Benestad, Rasmus E., Teichmann, Claas, Christensen, Ole B., Katharina, Bülow, Buonomo, Erasmo, Katragkou, Eleni, Steger, Christian, Sørland, Silje, Nikulin, Grigory, McSweeney, Carol, Dobler, Andreas, Palmer, Tamzin, … Brands, Swen. (2023). *EURO-CORDEX CMIP6 GCM Selection and Ensemble Design: Best Practices and Recommendations*. Zenodo. https://doi.org/10.5281/ZENODO.7673400

Taranu, I. S., Somot, S., Alias, A., Boé, J., & Delire, C. (2022). Mechanisms behind large-scale inconsistencies between regional and global climate model-based projections over Europe. *Climate Dynamics*. https://doi.org/10.1007/s00382-022-06540-6

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2053–2075. https://doi.org/10.1098/rsta.2007.2076

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., & Knutti, R. (2020). Past warming trend constrains future warming in CMIP6 models. *Science Advances*, *6*(12), eaaz9549. https://doi.org/10.1126/sciadv.aaz9549

von Trentini, F., Leduc, M., & Ludwig, R. (2019). Assessing natural variability in RCM signals: Comparison of a multi model EURO-CORDEX ensemble with a 50-member single model large ensemble. *Climate Dynamics*, *53*(3), 1963–1979. https://doi.org/10.1007/s00382-019-04755-8

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0189.1.