Performance based sub-selection of CMIP6 models for impact assessments in Europe

Tamzin E. Palmer[1], Carol F. McSweeney[1], Ben B.B. Booth[1], Matthew D.K. Priestley[2], Paolo Davini[3], Lukas Brunner[4], Leonard Borchert[5], and Matthew. B. Menary[6]

[1]Met Office Hadley Centre, FitzRoy Rd, Exeter, Devon, EX1 3PB, UK
[2]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK
[3]Consiglio Nazionale delle Ricerchere, Istituto di Scienze dell'Atmosfera e del Clima (CNR-ISAC), Torino, Italy
[4]Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria
[5]Climate Statistics and Climate Extremes, Centre for Earth System Research and Sustainability (CEN), Universität Hamburg, Germany. Laboratoire de Météorologie Dynamique (LMD) at École Normale Supérieure (ENS), Paris, France
[6]Laboratoire de Météorologie Dynamique (LMD) at École Normale Supérieure (ENS), Paris, France

Correspondence: Tamzin Palmer (tamzin.palmer@metoffice.gov.uk)

Abstract. We have created a performance-based assessment of CMIP6 models for Europe that can be used to inform the sub-selection of models for this region. Our assessment covers criteria indicative of the ability of individual models to capture a range of large-scale processes that are important for the representation of present-day European climate. We use this study to provide examples of how this performance-based assessment may be applied to multi-model ensemble of CMIP6 models to a) filter the ensemble for performance against these climatological/ processed-based criteria and ~~,~~b) create a smaller sub-set of models based on performance, that also maintains model diversity and the filtered projection range as far as possible.

Filtering by excluding the least realistic models leads to higher sensitivity models remaining in the ensemble as an emergent consequence of the assessment. This results in both the 25th percentile and the median of the projected temperature range being shifted toward greater warming for the filtered set of models. We also weight the unfiltered ensemble against global trends. In contrast this shifts ~~both the distribution of~~ the distribution towards less warming. This highlights a tension for regional model selection in terms of selection based on regional climate processes versus the global mean warming trend.

1 Applications and motivations for regional sub-selection

Global Climate models (GCMs) represent one of the key datasets to explore potential future climate impact, vulnerabilities and risks. However, not all GCMs are equally skilful in capturing the climate processes that drive climate variability and change,

particularly at regional scales (Eyring et al., 2019). There is a growing interest, therefore, in assessing models and selecting them for their suitability, if they are to be used to underpin or inform decision making. Such assessments are time consuming, often pulling on diverse strands of evidence across the important physical and dynamical processes, which will vary according to region, application and variable of interest. This assessment information is also not commonly available to the broader public making or using climate projection information. In this study we illustrate how such an assessment can be made for the Coupled Model Intercomparison Project 6 (CMIP6) generation models, for projections in European regions. This provides an assessment of how well these current models are able to capture the important regional processes over Europe. This information can either be used by those focusing on particular processes or as a combined assessment, to identify which subset of models may be more able to capture the relevant drivers of European climate change.

Historically, the climate modelling community has been cautious about weighting or eliminating poorly performing members due to the difficulties of linking performance over the historical period with future projection plausibility defaulting to a 'one model, one vote' approach (e.g. Knutti, 2010; IPCC, 2007, 2013). ~~However, an~~ Whetton et al. (2007) evaluate the link between model performance in the historical period and model performance for future projections by investigating the model similarity in patterns of the current climate and the inter-model similarity in regional patterns in response to CO2 forcing. They find that similarity in current climate regional patterns of temperature, precipitation and MSLP from GCMs is related to similarity in the patterns of change of these variables in the models.

In addition while global temperature biases in the historical record are not correlated with future projected warming (e.g., Flato et al., 201 . This is not the case regionally for Europe, where biases in the summer temperatures have been found to be important for constraining future projections (Selten et al., 2020). In addition projections of the Artic sea ice extent have also been linked to historical temperature biases (Knutti et al., 2017).

An increasing body of literature ~~demonstrates cases where there are shortcomings~~ does link short comings in the ability of ~~an individual~~ a model to realistically represent an observed baseline ~~climatology are~~ to being an indicator that the ~~model's~~ models' future projections are less reliable ~~(e.g. Whetton et al., 2007; Overland et al., 2011; Lutz et al., 2016; Jin et al., 2020; Chen et al., 2022; Ru .Further, for the purposes of applying models in impact studies at the regional scale, very unrealistic models may be of limited use if they require excessive bias correction.~~ (e.g., Whetton et al., 2007; Overland et al., 2011; Lutz et al., 2016; Jin et al., 2020; Chen et al. .

Regional model sub-selection is guided by a range of choices and there is always an element of subjectivity in terms of how the criteria are determined. For example, if a model performs well for a particular target variable, but then performs poorly in another season, variable, or location, this indicates that the regional climate processes are suspect (Whetton et al., 2007; Overland et al., 201 .

To assess the model performance in terms of the regional climate processes, we firstly identify the key drivers of the European climate as our criteria. We then use these to assess the performance of the CMIP6 models across a range of variables. The approach that we take one of elimination rather than selection and we do not recommend any individual model. Rather in our examples of approach to sub-selection, we examine the impact on the projection range from the elimination of the models that perform relatively poorly in these key criteria.

While ~~these~~ there are strong arguments for filtering the ensembles for regional applications, the practical implementation requires us to navigate several challenges, such as how to select appropriate criteria, where the appropriate thresholds should lie for 'acceptable' vs 'unacceptable' models, and how to deal with models that perform well against some criteria but poorly against others. This inevitably introduces a degree of subjectivity in both the selection of the qualifying criteria and deciding the appropriate thresholds. For example, assessments of future changes in ~~wintertime~~ winter-time extreme rainfall in northern Europe are likely to emphasise the ability of simulations to capture the observed storm track position, whereas those assessments looking at summertime heat waves in central Europe may place more emphasis on ability of models to adequately represent summer blocking and land-atmosphere interaction processes. Advances in model development have led to significant improvements in the realism of regional processes, with incremental improvements in a number of long-standing biases and key processes (Bock et al., 2020).

~~The assessment~~ Assessments and sub-selection of GCMs for regional applications ~~has been attempted~~ have been implemented for CMIP6 using metric based approaches (e.g. Zhang et al., 2022; Shiogama et al., 2021). These studies aim to score or rank models for a particular region (Shiogama et al., 2021) or a range of regions based on a number of metrics (Zhang et al., 2022). Other regional approaches may weight GCMs based on regional performance against a range of metrics (e.g. Brunner et al., 2019). Weighting models regionally based on a range of metrics may produce mixed results however, and not always improve the ensemble mean bias. Assessments that are based on process-based analysis that emphasise region-specific process may produce better results (MS and JA, 2019). The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP), aims to collate climate impact data that is consistent for both global and regional scales and across different sectors (Rosenzweig et al., 2017; Lange and Büchner, 2021). These studies use a limited number of GCMs (from CMIP5 and CMIP6) that are largely selected based on the availability of daily data for the required variables (Hempel et al., 2013). There have been concerns however, that the 4 GCMs used from CMIP5 in ISIMIP2b ~~may be~~ maybe unable represent the full range of uncertainty for future climate projections, especially for precipitation (McSweeney and Jones, 2016; Ito et al., 2020).

In this paper, we illustrate how current climate models can be assessed against their ability to capture a broad range of ~~large scale~~ large-scale climate processes important for the European climate in the recent historical period~~(1995-2014)~~. The rationale for doing so, is that models which do not adequately represent processes known to be important in the historical period for Europe, will not provide useful projections of future changes in these processes.

A processed based assessment, such as this, has several useful potential applications:

1. More robust European climate projections. By excluding models with the least realistic representation of regional climate drivers, we ensure that European projections are based on only those which can adequately capture present day processes. These remaining models are better candidates for understanding downstream impacts, both because their model biases are likely to be reduced compared to models that are unable to represent key features of the climate in the historical period, and because we can have more confidence that they can capture the regional processes relevant to future changes.

2. Assess whether ~~process based~~ process-based evaluation has impact on the range of expected future changes. Such an assessment provides an opportunity to explore whether there may be any relationship between the quality of regional process representation and the range of changes projected from these models.

3. As an aid to further model development. Identifying where individual climate models have problems with particular regional climate processes, can be used to inform the type of model processes where further model development would be beneficial, both for individual models and for GCMs in general.

4. Define a reduced set of more reliable climate projections to inform subsequent sub-selections. Several approaches make use of small(er) subsets of simulations, for computational or practical reasons or to simplify climate projection information. A performance filtered subset ensemble represents an important starting point for such ~~as~~ a selection and there are different approaches that may be used:

a. Sub-selection matrix: Sub-selection is often used to identify a simpler set of data that retains the characteristics of the underlying range of projected changes. This might be motivated either by computation (or other practical) limitations on the number of models and /or climate realisations that can be used in a particular application. In the case of sub-selecting a GCM matric for downscaling, Regional Climate Models (RCMs) will inherit errors from GCM boundary conditions. Therefore selection of models based on their ability to reproduce regional boundary conditions, such as features of large scale circulation is desirable (MS and JA, 2019). Alternatively, it might be motivated by the desire to reduce the complexity, by sub-selecting from the multi-model ensemble to still represent the underlying distribution as far as possible. Here, there is a need to balance criteria on credibility, with criteria to ensure that the subset can capture the broader range of potential changes and consists of as many independent models as possible.

b. Selecting individual realisations for use as climate narratives: Individual realisations are often used to exemplify responses in certain parts of potential climate projection space. For example, selecting realisations to represent what central estimates or ~~worst case~~ worst-case estimates, of future changes might look like. Alternatively selecting realisations that can be used to illustrate changes by particular drivers (e.g. the impact of strong changes in the NAO van den Hurk et al., 2014)

or dynamical drivers of regional changes (e.g. Shepherd, 2019, 2014; Zappa and Shepherd, 2017). Pre-filtered ensembles based on regional performance metrics help identify more credible realisations that could be used as climate narratives.

Here we demonstrate performance filtering for CMIP6 models, against a broad range of climate ~~process~~ process-based criteria relevant to Europe. This filtered subset can be used as a starting point, by others, to inform a selection of climate simulations appropriate for their own applications. This could be used by either drawing on individual assessment criteria or, as we go on to show here, the outcome of filtering on the full set of assessment criteria. In this paper, we illustrate the implication of this filtering for the range of expected changes over Europe (point 2, above) and work through an example of how this could be used in conjunction with model diversity criteria, to identify a smaller subset of realisations suitable for driving downstream impacts relevant modelling.

The selection of GCMs for a particular region is an opportunity to exclude models that are considered to be ~~inadequate~~ 'Inadequate' in terms of their ability to represent key drivers of the regional climate. This has been attempted in a number of studies (McSweeney et al., 2015; Lutz et al., 2016; Prein et al., 2019; Ruane and McDermid, 2017), but it is still a challenge in terms of how to identify which models are ~~inadequate~~ 'Inadequate' and how the decision to eliminate these models should be made, particularly if their removal results in a significantly reduced projection range. Where the removal of a model, that is not considered to be able to give meaningful or useful information about the present or future climate, reduces the range of projections, this needs to be carefully justified. In addition to classifying models as either adequate or ~~inadequate~~ 'Inadequate', we look to classify models in a more informative way, and provide further information about how each of the CMIP6 models may perform in terms of key processes that influence the climate in the main European regions. The assessment is broken down into a number of different criteria that are scored individually, providing information regarding how individual models perform for each of these.

We build on the approach developed in McSweeney et al. (2015, 2018) previously applied to CMIP5. In McSweeney et al. (2015, 2018), CMIP5 models were assessed on a range of regional criteria, including the circulation climatology, distribution of daily storm track position, the annual cycle of local precipitation and temperature in European sub-regions. These characteristics were assessed using a qualitative framework for flagging poorly performing models as 'implausible', 'significantly biased' or 'biased'. This performance information was subsequently used together with information about projection spread (McSweeney et al., 2015) or model inter-dependencies (McSweeney et al., 2018) to arrive at ~~subsets~~ sub-sets of the required size.

Many of the individual models and higher resolution model versions in CMIP6 show significant improvements in common model biases compared to CMIP5 (Bock et al., 2020). There are also a number of assessments in the literature that show an improvement in many of the processes that are key drivers of the climate for Europe e.g., Storm Tracks ~~(Priestley et al., 2020, 2022)~~ (Priestley et al., 2020, 2023), Blocking frequency (Davini and d'Andrea, 2020) and North Atlantic (NA) Subpolar Gyre (SPG)

sea surface temperature (SST) (Borchert et al., 2021b). We draw on these analyses already in the literature to assess these large scale large-scale processes for the European region, along with the assessment of features such as large scale circulation patterns, precipitation annual cycle and surface temperature biases using the method of McSweeney et al. (2015). Additionally, we look to classify models in a more informative way than simply keep or reject for sub-selection, to provide further information about how that model may perform in terms of key processes that influence the climate in a particular European region. Finally we note that our assessment is based solely on process base process-based criteria and does not use and any regional or global warming trends, which separates it from many recent global assessments of CMIP6 (Tokarska et al., 2020; Brunner et al., 2020b).

In the following section we describe how each of the classifications that we use for the criteria are defined. This is followed by a description of the criteria criteria that have been selected along with their relevance for the European climate. In section 3a) We then define how each of the classifications that we use for the criteria are defined. In section 3 we present the results of the assessmentmethodology along with examples of how individual criteria have been assessed. In section 3b) 4 we examine the impact of filtering out models that fail to reproduce key processes on the projected range. In section 3b) we We then use these performance filtered models to create a smaller sub-selection that also considers model diversity and maintains the projected range of the filtered models as far as possible. In sections 4 and 5 and 6 respectively, we discuss these results and present our conclusions.

2 Performance assessment for Europe

2.1 Classification definitions

The purpose of this assessment is to identify models within the multi-model ensemble that are less capable of reproducing the climate processes that are relevant for the regional European climate. In terms of assessing the plausibility and performance of climate models, a degree of subjectivity is involved. A mix of quantitative (RMSE, bias, variance, correlation) and qualitative (e.g. circulation wind patterns, SPG gradients) have been used and graded using coloured flag system. This approach has been chosen, as opposed to a more quantitative metric for the assessment, to indicate the model performance. While a numerical approach is in the first instance more objective, there is still the difficulty of determining the implications of a given error or bias when making decisions to include or exclude models. In our assessment, models are therefore grouped into classifications that we define in this section.

Red: Inadequate in performance criterion and should therefore be excluded from the sub-sample.

Orange: Unsatisfactory, substantial errors in remote regions where downstream effects could be expected to impact on the reliability of regional information and/or present in the local region of interest.

White: Satisfactory, model errors not widespread or not substantial in the local region of interest. Location of substantial remote errors are not known to have a downstream impact in the local region of interest. Captures key characteristics of the criteria spatially or temporarily.

~~Grey: Data/ analysis not available.~~

2.1 Criteria

2.1.1 Atmospheric criteria

The near surface temperature and precipitation are key variables for future climate and are of primary consideration in impact studies, especially in terms of future hydrology considerations (e.g. White et al., 2011; McDermid et al., 2014; Ruane et al., 2014). They have been considered key variables in previous sub-sampling approaches (e.g. Ruane and McDermid, 2017; McSweeney et al., 2015).

A number of previous studies have considered the importance of capturing the main synoptic features and large-scale atmospheric circulation patterns (e.g. McSweeney et al., 2012, 2015; Prein et al., 2019) as a key criteria for GCM sub-setting. For northern Europe in particular large scale weather patterns and the passage of weather systems that make up the North Atlantic (NA) storm track dominate the climate, especially in the winter. Extratropical cyclones are the dominant weather type in mid-latitudes where they can have a significant impact due to associated extreme precipitation and windspeeds (Browning, 2004; Priestley et al., 2020). They have an important role in the general circulation in the poleward transport of heat, moisture and momentum (Kaspi and Schneider, 2013) and in maintaining the latitude westerly flow. In the winter (DJF) many GCMs have a southern bias in the peak storm track density with the prevailing winds too zonal resulting in higher than observed windspeeds across central Europe (Priestley et al., 2020; Zappa et al., 2013). In the summer (JJA) the prevailing wind direction is more westerly and less strong, but still an important driver of weather systems and key for representing the climate. We assess the large-scale circulation by comparing a baseline climatology with the ERA5 data (e.g. 1995-2014), using a similar approach to McSweeney et al. (2015). We use the analysis of Priestley et al. (2020) to assess the NA storm track over Europe in individual CMIP6 models.

Blocking by high pressure weather systems is known to cause of periods of cold dry weather in the winter and summer heatwaves. Blocking is typically under-represented in GCMs and this is still the case large parts of Europe in CMIP6, although there has been some improvement in the bias in many CMIP6 models (Davini and d'Andrea, 2020; Schiemann et al., 2020). We use the results of the analysis carried out by Davini and d'Andrea (2020) to assess the performance of the CMIP6 models based on RMSE, bias and correlation.

2.1.2 Ocean criteria

The literature indicates that there is a link between NA Sea Surface Temperature (SST) and variability in the European climate (e.g. Dong et al., 2013; Ossó et al., 2020; Carvalho-Oliveira et al., 2021; Börgel et al., 2022; Sutton and Dong, 2012; Booth et al., 2012; Borchert et al., 2021a). The link between NA SST and drivers of the European climate is complex and how the

atmosphere and NA interact over different timescales has not been fully determined.

Representation of the NA SSTs in GCMs has also been shown to be key for other features such as blocking frequency ~~(Scaife et al., 2011; Keeley et al., 2012) (Sutton and Shaffrey, 2012), Storm Tracks (Priestley et al., 2022)~~ (Scaife et al., 2011; Keeley et al. , Storm Tracks (Priestley et al., 2023) and the NA jet stream (Simpson et al., 2018). GCMs commonly feature a cold bias to the south of Greenland Tsujino et al. (2020), which is associated with biases in the latitude of the North Atlantic storm track due to unrepresented latent heat fluxes Priestley et al. (2023)). This cold bias commonly causes the storm track to be situated too far south Athanasiadis et al. (2022). Removing this SST bias results in improvements in the latitude of the atmospheric circulation Keeley et al. (2012) and in the simulation of other atmospheric phenomena such as blocking Scaife et al. (2011)).

If this link between NA SSTs and the European climate remains important in the future a ~~satisfactory~~ 'Satisfactory' representation of NA SST is required for also predicting the future European climate ~~(e.g.)Gervais2019, Oudar2020~~(e.g. Gervais et al., 2019; Oudar . There also appears to be some improvement in the skill in representation of the decadal NA and Sub-polar gyre in particular in CMIP6 compared to CMIP5 (Borchert et al., 2021b), which may be a factor for improvements in the representation of storm tracks (Lee et al., 2018), and blocking frequency (Keeley et al., 2012) for the European region in CMIP6 models compared to CMIP5.

The AMOC also plays a significant role in the present and future European climate due to its role in the poleward transfer of heat and ocean circulation. It also impacts on the NA SST (Jackson et al., 2022; Zhang, 2008; Zhang et al., 2019; Yeager and Robson, 2017), thereby influencing the SST impact on European climate discussed above. The CMIP5 and CMIP6 ensemble both predict a reduction in the AMOC by the end of century for higher emissions pathways, (Menary et al., 2020; Bellomo et al., 2021). The AMOC model comparison with rapid data from the analysis of ~~.~~(Menary et al., 2020) is used to assess the AMOC the GCMs.

2.2 Classification definitions

The purpose of this assessment is to identify models within the multi-model ensemble that are less capable of reproducing the processes that are relevant for the regional European climate. In terms of assessing the plausibility and performance of climate models, a degree of subjectivity is inevitably involved. One approach is to assess and rank the performance of the models based on a number of purely numerical measures of model error (RMSE, bias, variance, correlation), this provides valuable and objective information about the relative performance of the models, but it does not assess what the implications of the errors are, in terms how they impact on the ability of the model to make a meaningful regional projection. An additional qualitative element to the assessment can add value by interpreting how these errors impact on the overall performance of the model for the regional climate and helps to inform the question of why these errors may cause a model projection to be less reliable.

A mix of quantitative (RMSE, bias, variance, correlation) and qualitative (e.g., inspection of circulation wind patterns) have been used and the models graded for each criterion using a coloured flag system. Visual inspection allows us to understand the characteristic of the error and consider its impact on other aspects of the model.

The models are given a classification 'flag' for each of the criteria described in the previous section, creating a table or coloured map summarising the performance of each model. This approach has been chosen, as opposed to a more quantitative metric for the assessment, to indicate where model performance for a variable is an issue. Where the qualitative assessment has been applied the quantitative metrics have been used as a guide to sort the models into classifications and also to ensure consistency as far as possible. The the full details of how this has been applied to each criterion are described in the appendices (two examples are also given in the following section) In our assessment, models are therefore grouped into classifications that we define in this section.

Red: 'Inadequate' in performance criterion and should therefore be excluded from the sub-sample.

Orange: "Unsatisfactory" substantial errors in remote regions where downstream effects could be expected to impact on the reliability of regional information and/or present in the local region of interest.

White: 'Satisfactory', model errors not widespread or not substantial in the local region of interest. Location of substantial remote errors are not known to have a downstream impact on the local region of interest. Captures the key characteristics of the criteria spatially or temporarily.

Grey: Data/ analysis not available.

3 Materials and Methods

3.1 Data sources

Details of models from CMIP6 multi-model ensemble (Eyring et al., 2016) that are included in this study can be viewed in table ~~SM1~~ S1 in the supplementary ~~material~~information.

We use a baseline period of 1995-2014 and ~~a future periods~~the period: 2081-2100 (end of century) for future projections. These time periods have been selected for consistency with existing EUCP analyses ~~(e. g. Brunner et al. (2020a)).~~ (e.g., Brunner et al., 2020a). We use the SSP585 scenario for comparison as this is the scenario with the strongest climate signal. The model data for the large area averages (for comparison of temperature and precipitation changes) is regridded onto a 2.5° x 2.5° grid and land-sea mask applied as used in ~~(Brunner et al. (2020a); Palmer et al. (2021))~~Brunner et al. (2020a) and Palmer et al. (2021), using a standard nearest neighbour interpolation. The data was averaged spatially using a weighted area

mean.

~~Reanalysis and observational data along with details of the historical time periods that were used for the assessments of the individual criteria, in addition to examples~~ The ERA5 reanalysis data and E-OBS gridded observational dataset (to evaluate the precipitation annual cycle) were used to assess the model error. Monthly mean data is used for the assessment with the exception of the blocking frequency analysis which uses daily data fields. Details of how these assessments have been carried out for each of the criteria, are given in the appendices. Examples of the assessments are also shown in the following section (section 3.2) for large scale circulation and storm tracks~~are given in the appendices~~.

We use the results of the assessment, as described in the previous sections and summarised for the CMIP6 models, where sufficient assessment information was available in ~~Figures 5and ??. Details of how the assessment was made for each of the criteria are given in appendix.~~ Figure 5.

We use only the first realisation for each of the models in this assessment and assume that this is generally representative of the model performance. We acknowledge however that there may be a role for internal variability that pushes a model across assessment classifications. The largest uncertainty due to internal variability of the diagnostics we use is likely to be from the historical trends (which are not part of the assessment but used in an illustrative capacity). Brunner et al. (2020) found that for the global case the spread in the temperature trend fields between ensembles members of one model can be in the same order of magnitude as the spread across the multi-model ensemble. For the temperature climatology, in turn, the spread between ensemble members of the same model is typically less than 10% of the multi-model spread. This gives some indication that we can expect there to be relatively low variation in the performance of the models across the climatology for temperature based on which member is used. For the AMOC, which is a significant contributor to regional and global climate variability, Menary et al. (2020) noted that links to North Atlantic SSTs were sensitive to the removal (or not) of forced variability, but individual model realisations were not systematically different.

A case study is made to assess the role of internal variability for large-scale circulation (in which we may expect larger variability across ensemble members, than for the temperature climatology) in the CanESM5 model across all 25 realisations. This can be viewed in the supplementary information (Figs. S5 and S6). This context suggests that the analysis presented in this paper, based on the first ensemble member, likely provides an indicative picture typical of the response across any wider initial condition ensemble. However, future assessments may want to look for individual ensemble members which may show weaker manifestations of particular biases, particularly where a model lies close to classification boundaries.

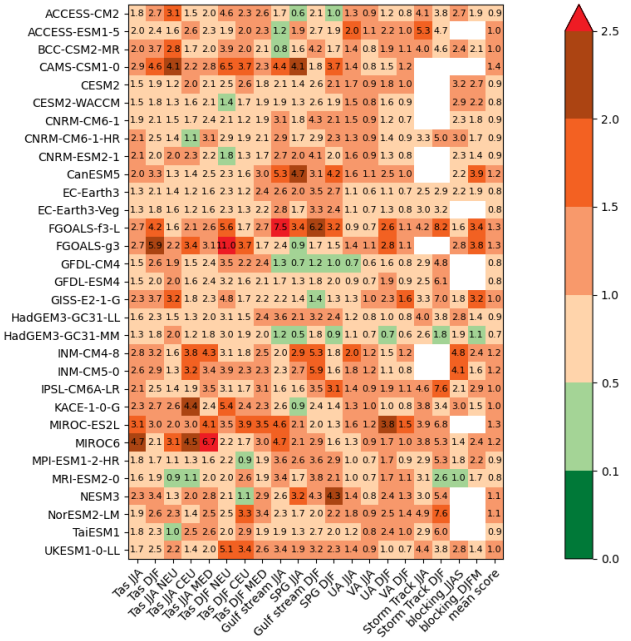


Figure 1. Summary of RMSE values for the large-scale assessment criteria. The colour scale is determine by the ratio of the model RMSE to the ensemble mean RMSE. . RMSE values are absolute, the mean score is the average of the relative error (normalised by the ensemble mean) across each of the criteria.

3.2 Assessment examples

In this section we show two examples of the assessment method for two of the criteria discussed in section 2. Examples for all the criteria are given in the appendices.

For the assessment of each criterion, we refer to the model RMSE, bias and in some cases correlation with the reanalysis (e.g., for the precipitation annual cycle, see appendix A1 for details), in addition to a qualitative assessment of the model climatology, in terms of how errors impact on the ability of the model to represent the regional climate. In the process of classifying the performance of the models, the qualitative interpretation of the errors has an element of subjectivity, as does the decision of where to place various thresholds for the quantitative measures. We aim to keep the assessment process as transparent as possible. In addition, it is important that while the qualitative assessment for an individual classification may occasionally differ to some degree from a purely quantitative approach, these decisions should not lead to the retention of models with objectively larger errors in the sub-selection process. In the following section 3.2.1 (and in the appendices) we refer to both the fields of the model climatology and Fig. 1, which summaries the RMSE for each of the models.

3.2.1 Large scale circulation patterns

The ~~large scale~~ large-scale seasonal circulation pattern was assessed for winter (DJF) and summer (JJA) based on the mean climatology at 850hPa for the baseline time period 1995-2014, the ERA 5 reanalysis was used for comparison (Fig. 2).

In DJF, European weather is dominated by the passage of weather systems that make up the NA storm track, the prevailing direction for these is from the south-west as can be seen in the climatology in ERA5 (Fig 2a). ~~This pattern is most evident in the models classified as satisfactory (e. g. CNRN-CM6-1 and CESM2). The strength of the south-westerlies over the UK and Scandinavia is too weak in some of the models (e. g. BCC-ESM1 Fig . 2f) , along with the prevailing wind direction being too zonal, which leads to a positive bias for wind-speed in central Europe. These models were flagged as 'unsatisfactory'.The largest errors for the satisfactory~~ The model large-scale RMSE for the 850hPa wind vectors along with a qualitative assessment of the overall circulation pattern were used to assess the models for this criterion. Fig. 1 show that the wind vectors RMSE is less that of the multi-model mean for CNRM-CM6-1 and HadGEM-GC31-LL. Where the wind vector errors for a model is less than the multi-model mean the large-scale circulation is found to be reasonably well represented. Fig 2b) and c) show that these models capture the overall circulation pattern well and have relatively low windspeed biases. Where the models have a larger RMSE for wind vectors than the multi-model mean, the threshold for an 'Unsatisfactory' model requires some consideration. For these cases a qualitative approach is used understand how these errors may impact on the European climate and to guide where this threshold should lie.

The model with the largest errors of the 'Satisfactory' models is CESM2, with an area of positive bias over the UK, this model was still assessed as ~~satisfactory~~ 'Satisfactory' however, due to the well define ~~southwesterly~~ south-westerly wind patterns and good representation of the winds over most of the European land areas.~~The unsatisfactory models~~

The strength of the south-westerlies over the UK and Scandinavia is too weak in some of the models (e.g. IPSL-CM6A-LR Fig. 2f), along with the prevailing wind direction being too westerly. These models were flagged as 'Unsatisfactory'. These models feature a variety of structural biases, for example INM-CM4-8 which had a ~~similar~~ lower spatially averaged RMSE ~~error~~ windspeed error, but lacked a clear representation of the south westerlies over northern Europe~~, with winds~~ . The winds are too weak in these areas and ~~some~~ there are areas of negative bias in the Mediterranean. This model was classified as ~~unsatisfactory~~ 'Unsatisfactory' due to its lack of representation of the circulation pattern ~~(Fig~~ and general wind direction too westerly (Fig. 2g). This is also reflected in the wind vector errors (Fig. 1).

~~For the unsatisfactory models there was generally strong positive bias over European land regions, due to wind patterns that were too westerly. The BCC-ESM1 model captures some of the southwesterly wind pattern but has some areas of substantial bias (> 5m/s) over the Mediterranean and NA. The GISS-E2-1-G model has a similar but more pronounced pattern of errors .~~

~~The latter is classified as unsatisfactory rather than excluded despite a relatively large magnitude of errors (RMSE 2.39 m/s) due to the model capturing some representation of the circulation pattern (Fig2)~~

Models ~~selected for exclusion (flagged Inadequate)~~ flagged 'Inadequate', have an almost entirely westerly (no south-westerlies) wind pattern and the wind speed errors over large parts of Europe are widespread and substantial (e.g.~~Fig~~, Fig. 2h-j, CanESM5, FGOALS-g3~~, CESM2-FV2~~). These models ~~had~~ nearly all have a large (positive) bias over European land regions (e.g. > ~~6m/s).~~ $6ms^{-1}$). MIROC-ES2L has the largest errors for the wind vectors in the ensemble for DJF (more than twice the ensemble mean error for UA), the errors do not follow the same pattern as the other 'Inadequate' models, with a large weak bias over most of Europe and an almost northerly wind direction in the NA (Fig.2i).

Circulation patterns are more westerly with weaker winds in the summer (JJA). These were assessed using the same approach for comparison as for winter circulation (Fig 3). Many CMIP6 models capture the general pattern well (e.g HadGEM-GC31-LL, GFDL-ESM4, Fig 3b and d)~~, although some~~ . The UA, VA RMSE for both of these models is less than the ensemble mean RMSE. Again, where the models perform above the average for the multi-model ensemble the overall circulation pattern is well represented with relatively low windspeed bias.

As with the case for the DJF circulation where the models have larger errors than the multi-model mean for JJA wind vectors, the threshold is to warrant a flag as 'Unsatisfactory' or 'Inadequate' as been determined alongside some qualitative interpretation of the model errors.

Some of the models had westerly patterns over the UK and central Europe that were too weak (e.g. MIROC6, INM-CM4-8, Fig. 3~~e and g), leads to~~ f and h), as a result there are larger errors in European land regions and these models were therefore classified as ~~unsatisfactory. In~~ 'Unsatisfactory' or in the case of INM-CM4-8 where these errors are more pronounced , 'Inadequate'. In the case of MIROC6 we note that the magnitude of the UA and VA errors over the large-scale region assessed as a whole are on the borderline of the threshold for 'Satisfactory' and 'Unsatisfactory' compared to the other models. It is the relatively weak circulation and low bias in windspeed over the European land regions that is the reason for the 'Unsatisfactory' flag in this case (Fig. 3e).

The INM-CM4-8 (and to a ~~few cases the representation of the circulation pattern was considered poor enough to warrant exclusion of a model (e. g.~~ similar extent the INM-CM5-0) model has some of the largest errors for the JJA wind vectors in the multi-model ensemble. It is noted that these models are also flagged as 'Inadequate' for both severe JJA blocking errors and severe errors in representing the annual precipitation cycle in central Europe. There are also issues with the temperature bias in central Europe for this model (flagged 'Inadequate'). These severe errors in central Europe are likely to be related to the representation of the large-scale circulation.

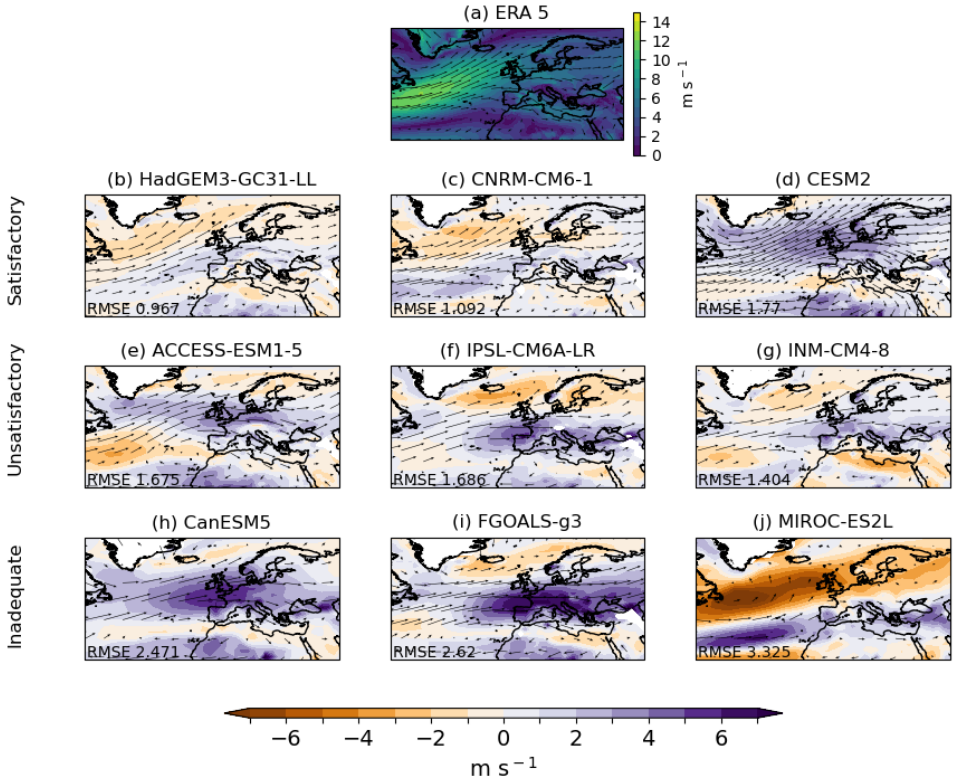


Figure 2. Examples of DJF circulation (850hPa) classifications for a sample of individual models. Top panel shows ERA5 climatology. Windspeed and direction are shown as a 20 year mean 1995 – 2014. Arrows show direction (absolute) of windspeed (scaled by windspeed) for climatology across all panels. The shading for the 3 panels ~~show~~ shows the difference in windspeed between the model and ERA climatology.

For NorESM2-LM ~~,~~ and ACCESS-ESM1-5) Fig 3i ~~) where~~ and j) the westerly pattern was too far north leading to a large area of positive bias over northern Europe. These models have the largest RMSE for wind vectors in the multi-model ensemble along with the INM-CM4-8 and INM-CM5-0 models and the largest RMSE for windspeed. The large region of substantial positive bias over the NA and much of Europe indicates that this error is likely to impact on the JJA storm track over Europe for these models. As the storm track assessment is available for both these models this can be confirmed to be the case. The storm track RMSE is in the top 85th percentile for the models assessed for the storm tracks, (see the following section on the storm track assessment) and Fig. 1 shows that these models have the largest errors for the JJA storm track in the ensemble.

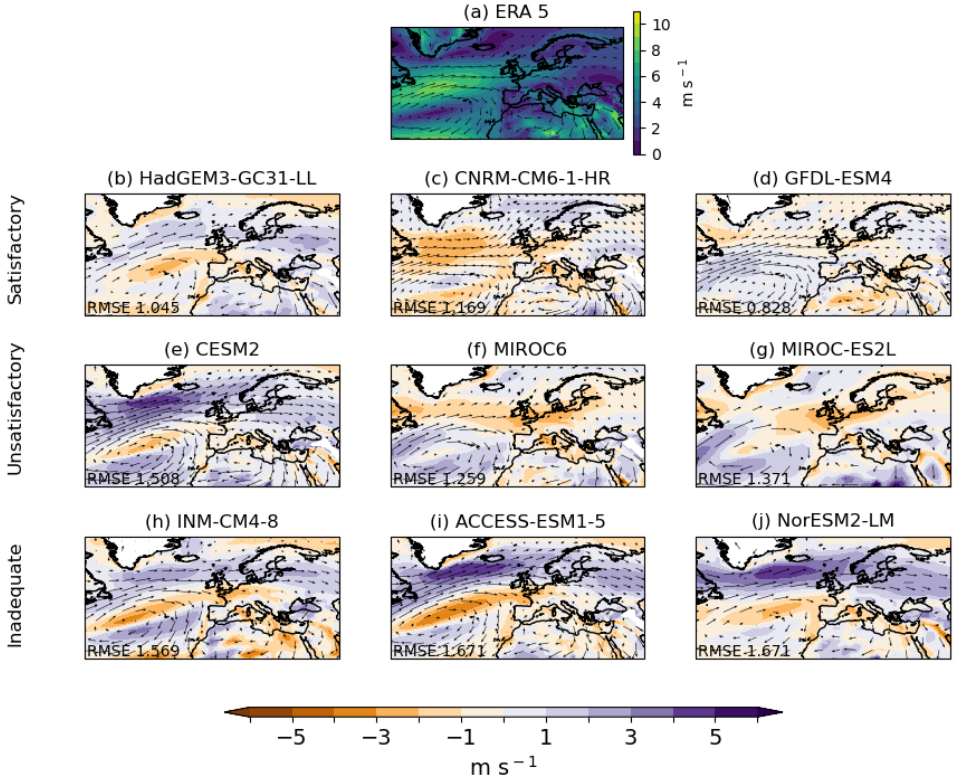


Figure 3. Examples of JJA circulation (850hPa) classifications for a sample of individual models. Top panel shows ERA5 climatology. Windspeed and direction are shown as a 20 year mean 1995 – 2014. Arrows show direction (absolute) and windspeed (scaled by windspeed) for climatology across all panels. The shading for the 3 panels show the difference in windspeed between the model and ERA climatology.

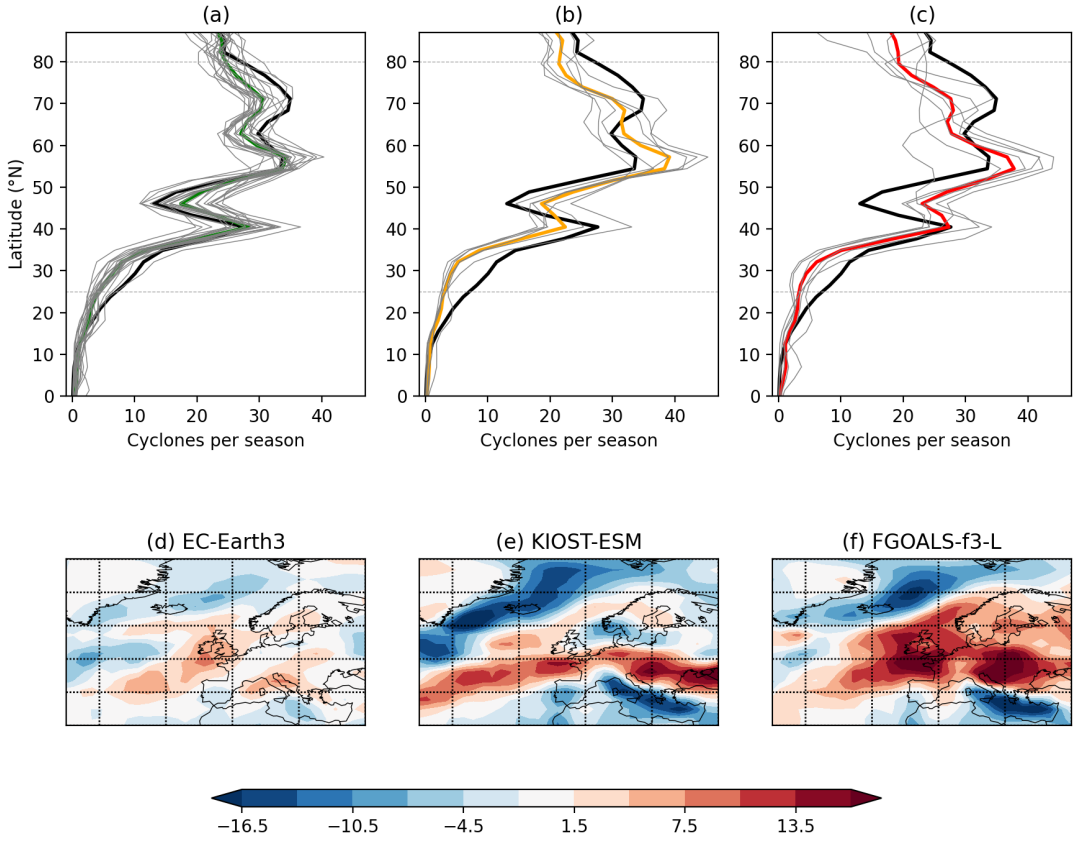


Figure 4. Examples of DJF storm track classifications. a), b) and c) shows the RMSE of the zonal mean track ~~20W -20E~~ 20°W -20°E for individual models and the classification mean, for 'Satisfactory' (a), 'Unsatisfactory' (b) and 'Inadequate' (c). In (a–c) gray lines are individual models, ~~solid coloured~~ solid-coloured lines are the group average, and black solid line is ERA5. Individual examples are shown in the lower panel for track density bias for 'Satisfactory'(d), 'Unsatisfactory' (e) and 'Inadequate' (f) models. Units of (d–f) are cyclones per season per 5 degree spherical cap.

3.2.2 Storm track large scale assessment

The ~~storm~~ track density is calculated using an objective cyclone tracking and identification method based on the 850 hPa relative vorticity (Hodges, 1994, 1995). The method and data are the same used in Priestley et al. (2020). The zonal mean of the model mean track density from 20°W-20°E was taken to get a profile of storm number by latitude. Then the RMSE was calculated of the models compared to the profile obtained from ERA5. The RMSE was calculated from 25-80°N.

The storm track has been assessed as a large scale feature using an assessment of the characteristic trimodal pattern (Fig 4) calculated as the zonal mean of the seasonal track density between ~~25N-80N and 20W-20E~~ 25°N-80°N and 20°W-20°E,

compared to ERA5 reanalysis data. The baseline time period used for this assessment is 1979/80-2013 (as in Priestley et al. (2020))~~and the model data is compared to ERA5~~.

The RMSE of zonal mean track density from ~~20W-20E~~ 20°W-20°E is used to initially sort the models into categories, while a hard cut off threshold was not applied for each category ~~the data~~ it was helpful to sort the models in < 65th RMSE percentile, 65th- 85th and > 85th percentile for RMSE. The different model groups were then inspected visually, it was found that although some of the models in the < 65th percentile had some significant biases the models in this group had clearly define peaks in their number of cyclones at the correct latitude and therefore captured the passage of storms across western and central Europe ~~satisfactorily~~'Satisfactorily' (Fig. 4a). This was not found to be the case for the models in the 65th - 85th RMSE percentile where there was a lack of a northern peak, this indicates a zonal bias in these models, which is a characteristic bias in GCMs (Fig. 4b). These models were classed as ~~unsatisfactory~~ 'Unsatisfactory' although the errors were not large enough on visual inspection to class them as ~~Inadequate~~'Inadequate', with the exception of MIROC-ES2L.

Models with >85th percentile RMSE failed to capture the tri-modal pattern, and had large biases in the number of cyclones at each of the peaks (Fig. 4c). In particular there was a lack of northern peak and an amplification of the errors in this group, with a large zonal bias in the track density. These models were considered unable to represent this feature and were flagged as ~~Inadequate~~'Inadequate'. Examples of individual models for each of the groups are shown in Figure 4d-f. The RMSE values for each of the models in the multi-model ensemble are also shown in Fig. 1.

3.3 Weighting for performance against global trends and model independence with the climWIP method

We also compare our results with the Climate model Weighting by Independence and Performance (ClimWIP) method (Knutti et al., 2017; Lorenz et al., 2018; Brunner et al., 2019, 2020b; Merrifield et al., 2020) to assess differences between our process-based filtering and an assessment based on historical warming. ClimWIP combines model performance weighting based on one or more metrics with an assessment of model independence (i.e., overlaps in the models' source code or development history). Here we use an adaptation of the approach described in Brunner et al. (2020b) and publicly available via the ESMVal-Tool (https://docs.esmvaltool.org/en/latest/recipes/recipe_climwip.html). Performance weights are calculated based on global temperature trends compared to ERA5 in the period 1980-2014. Independence weights are based on global model output fields for temperature and sea-level pressure which have been shown to reliably identify model dependencies (Brunner et al., 2020b; Merrifield et al., 2020). Here we use ClimWIP in two setups: one only based on performance weights, and one only based on independence weights as detailed later.

4 Results: Assessment and applications for sub-selection

4.1 Assessment ~~Tables~~Table

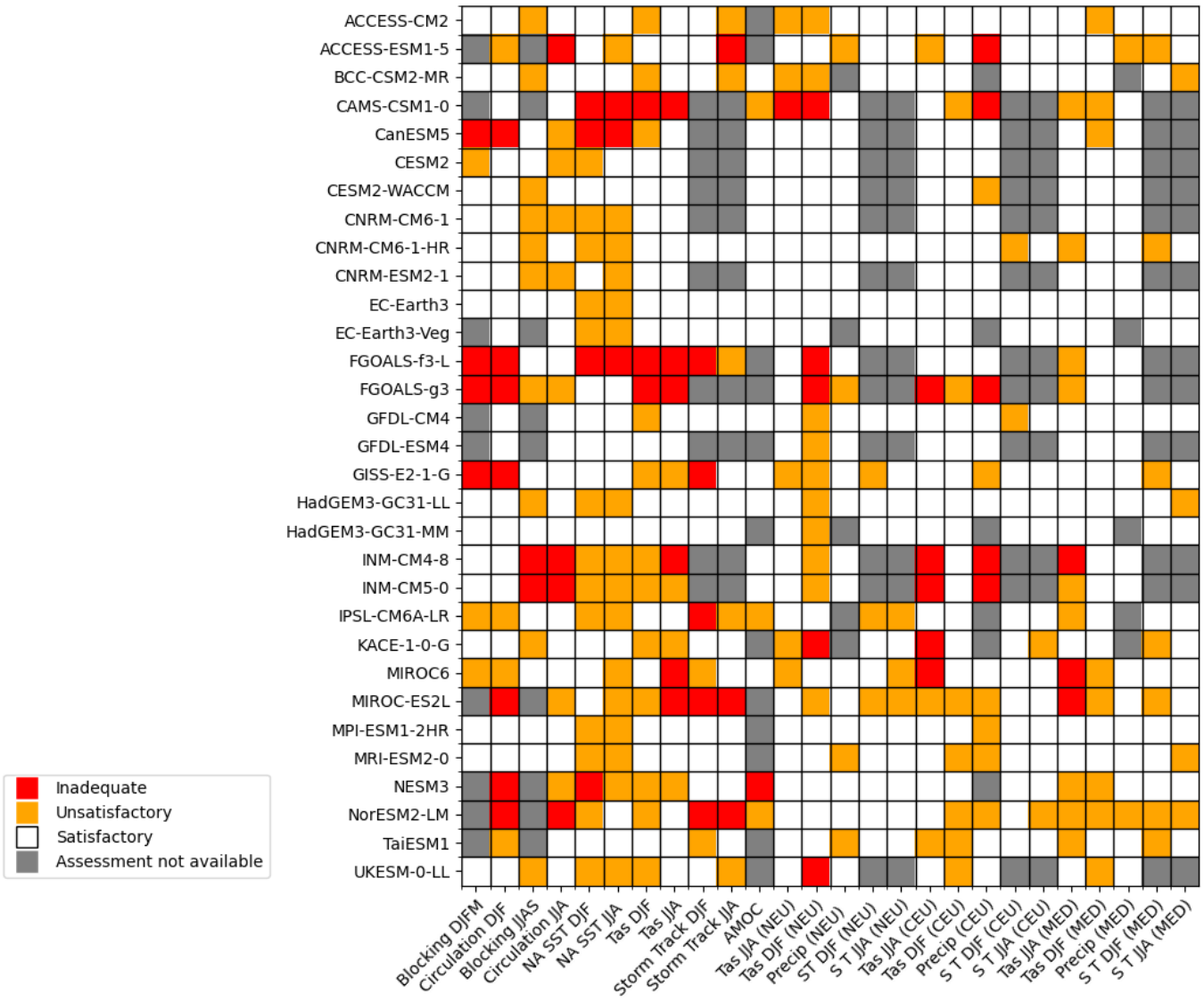


Figure 5. Model assessment summary for ~~large scale pan~~ qualitative European criteria. Assessment criteria for the large scale are as follows. Blocking: Blocking frequency, Circulation: large scale circulation assessed by 850hPa windspeed and direction, NA SST: NA SST bias, Tas: surface air temperature bias at 2m, Storm Track: based on RMSE of the zonal mean track ~~20W -20E~~20°W -20°E, AMOC: based on strength at 1000m at 26°N. Assessment criteria for the European regions are as follows. Tas: surface air temperature bias at 2m, Precip: Annual precipitation cycle, S T: storm track assessed as cyclones per season within European region.

~~Model assessment summary for regional criteria. Assessment criteria for the European regions are as follows.Tas: surface air temperature bias at 2m, Precip: Annual precipitation cycle, S T: storm track assessed as cyclones per season within European region.~~

The assessment for each of the CMIP6 models is collated into ~~two figures (~~Fig.5 ~~and Fig. ??)~~, with the classification for each of the criteria, where the relevant data and/or analysis are available. ~~The tables create~~ Fig. 5 creates a summary of each model's performance against a range of criteria, that are essential for a meaningful representation of the European climate.

~~Figure 5 and ?? summarise~~ This summarises the skill, across a multi-model ensemble from CMIP6, of their ability to capture ~~large scale processes and the presence or absence of large scale biases. In total, 15 models received a red flag for one or more criteria, and 21 models received 5 or more orange flags. Most models had at least 1 orange flag, but 15 were found have orange flags in less than 30% of the criteria they were assessed against. We find that models that show severe (red flag) issues in a particular criteria also produce issues in others (which is perhaps not surprising, given that they are not independent).~~the key process for the European climate.

The assessment criteria ~~is divided into two tables for large scale~~ are divided into large-scale and regional assessments. The ~~large scale~~ large-scale assessment criteria, such as ~~large scale~~ large-scale circulation and blocking frequency, are criteria that have a pan European ~~impacts~~ impact and are not specific to a particular region. The regional assessment criteria have been scored individually for each of the three main European regions used in the EUCP study and as defined as in Brunner et al. (2020a) and Gutiérrez et al. (2021). These are, Northern Europe (NEU), Western and Central Europe (CEU) and the Mediterranean (MED) (see Fig ~~SM1~~ S1, in the supplementary information). We focus in ~~this~~ the assessment on summer (JJA) and winter (DJF).

Some of the criteria were assessed ~~for~~ both at the large scale and regionally. For example, it is useful to know if a model has a widespread temperature bias that extends over Europe and the NA, but it is also the case that some models have more localised temperature ~~bias's~~ biases that affect individual regions. For the regional assessment where surface variable (e.g., precipitation, temperature) are assessed models were scored for their performance solely over the land regions.

The classifications in ~~Figs 5 and ??~~ Fig. 5 can be applied to create a bespoke sub-set of CMIP6 models depending to on the motivation for sub-selecting. Here we have used the red classification of ~~Inadequate~~ 'Inadequate' to indicate that a model should be removed, but it may be the case that a less strict approach to performance filtering is ~~required for some of the processes~~acceptable in some cases, than we have applied here. Likewise, it may be the case that an ~~Unsatisfactory~~ 'Unsatisfactory' (orange) flag in a certain ~~criteria~~ criterion such as the regional precipitation may be particularly undesirable. In the following section we use the table to create two different sub-sets from the multi-model ensemble.

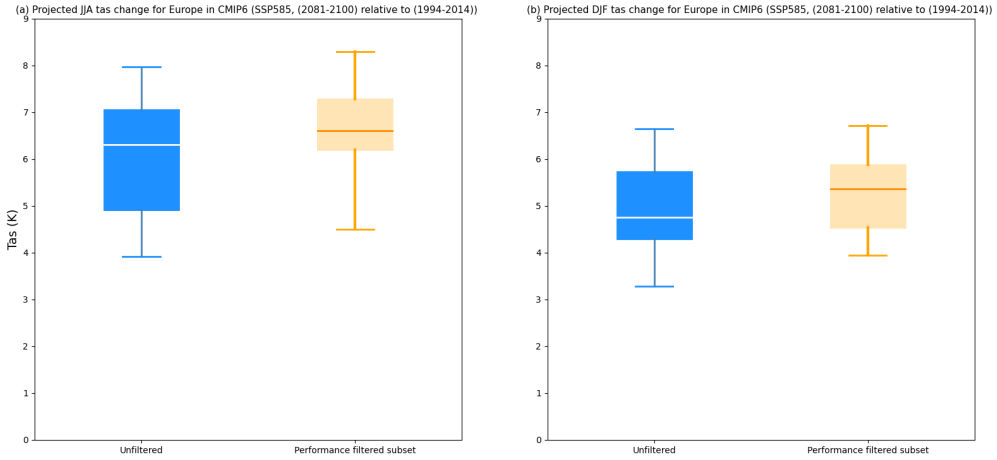
4.2 Excluding the models least representative of key regional processes

In this section we explore the implication of screening out poor ~~model variants~~ models based on the process-based performance assessment alone, on the range of projected regional changes~~(see point 2, in the introduction)~~. The aim is to revisit the range of projected regional climate changes, excluding those shown to struggle representing regionally relevant processes. ~~Here we exclude both models flagged as Inadequate(red flags, Section 3a), Fig . 5and **??**) as well models with Unsatisfactory performance in several criteria(orange flags) .~~ Note we do not include any criteria based on climate sensitivity or global temperature trends in this selection for this reason. These additional considerations and how they could be applied will be discussed along with the results.

~~We remove any models with with an Inadequate(red) flagin figure 5 or **??**. Remaining models with a substantial number of orange flags (we use a threshold of >/= 30%) are also removed. This leads to the additional removal of the UKESM-0-LL and TaiESM1 model. The TaiESM1 model has substantial errors (orange flags), across arange of large scale criteriasuch as storm tracks, NA SST and large scale circulation for DJF.It also has orange flags for temperature biases and precipitation in some regions~~

For the sub-selection process, we refer back to the definition of the classifications in section 2.2. The 'Inadequate' category (shown as a red flag on Fig 5) is used to indicate that a model fails to represent a key feature of the regional climate and should be removed from the sub-selection. We also differentiate between large-scale criteria than can be expected to have pan European effects on the model performance (and may also be inherited from the GCM in case of down-scaling) and regional criteria, which may only be of concern in the local region. Here we consider the impact on the projection range of firstly excluding any model with one or more 'Inadequate' (red) flags for any of the large-scale criteria. We then go on to consider any further changes in the projected temperature range as a result of removing an remaining models with a regional 'Inadequate' flag. ~~The UKESM-0-LL model has orange flags for JJA blocking frequency and storm tracks, it also has a widespread substantial winter temperature bias. The performance filtered subset is listed in table **??** along side the original raw unfiltered CMIP6 ensemble . To be used in the filtered ensemble each model is required to be assessed against a minimum of 50% of the assessed criteria, including at least one of the atmospheric criteria, NASST and temperature for the large scale criteria .~~

~~Models from CMIP6 ensemble that were assessed using Fig. 5 and **??** and the performance filtered models.Raw CMIP6 Performance filtered subsetACCESS-ESM1-5~~ Once all the models in Fig. 5 that have a red flag for the large-scale criteria are removed the following models remain in the sub-selection; ACCESS-CM2, BCC-CSM2-MR~~BCC-CSM2-MR CAMS-CSM1-0~~ , CESM2~~CESM2 ~~, CESM2-WACCM~~CESM2-WACCM ~~, CNRM-CM6-1~~CNRM-CM6-1 ~~, CNRM-CM6-1-HR~~CNRM-CM6-1-HR ~~, CNRM-ESM2-1~~CNRM-ESM2-1 EC-Earth3 CanESM5 ~~, EC-Earth3. EC-Earth3-Veg~~EC-Earth3 ~~, GFDL-CM4~~EC-Earth3-Veg ~~, GFDL-ESM4~~FGOALS-f3-L ~~, HadGEM3-GC31-LL~~FGOALS-g3 HadGEM3-GC31-MM GFDL-CM4 ~~, HadGEM-GC31-MM,~~ MPI-ESM1-2-HR~~GFDL-ESM4 ~~, MRI-ESM2-0, KACE-1-0-G, TaiEMS1, UKESM1-0-LL. This sub-selection from the qualitative assessment can also be compared to the RMSE values in Fig.1. If we look at the scores for the large-scale criteria (all categories



~~HadGEM3-GC31-LL~~

Figure 6. a) Projected range of JJA temperature change for Europe in CMIP6 (SSP585, (2081-2100) relative to (1994-2014)) for the raw unweighted multi-model ensemble and the large-scale performance filtered subset. Boxes show 25th to 75th percentile. Whiskers show the 5th and 95th percentile. b) As for a) but for DJF.

in Fig 1, excluding regional temperature), it can be seen that the excluded models include all those with a RMSE more than 1.5 times the ensemble mean in at least one of the large scale categories. It is also the case that for the retained models that the RMSE does not exceed 1.5 times the multi-model ensemble mean for any large-scale category. The retained models also perform better than, or at least equal to the ensemble mean across all the categories. This indicates that in our application of the assessment objectively poorer models have been removed (in terms of large-scale performance) and those with objectively smaller errors have been retained.

~~GISS-E2-1-G~~

Fig.6 shows the difference in the projected temperature range for the large-scale process-based filtered sub-set and the unfiltered multi-model ensemble. The difference in DJF is small (Fig. 6b), however in JJA the lower part of the range is reduced, and the upper part is shifted upwards (Fig. 6a). This shift in the projection range indicates that more of the higher sensitivity models are retained by filtering using process-based performance criteria.

~~HadGEM3-GC31-MM~~

In the second stage of filtering, we again refer to the regional criteria in the assessment table. There are 'Inadequate' (red) flags for regional precipitation (in central Europe) and for regional temperature in a few of the models (Fig.5). The models with an 'Inadequate' classification for precipitation (INM-CM4-8, INM-CM5-0, ACCESS-ESM1-5 and FGOALS-g3), already have at least one 'Inadequate' flag for the large-scale atmospheric criteria. Therefore, these models have already been removed from the performance filtered sub-set. The KACE-1-0-G ~~MIROC-ES2L MIROC6 MPI-ESM1-2-HR MRI-ESM2-0 NESM3~~

~~NorESM2-LM TaiESM1 UKESM1-0-LL~~ model has two 'Inadequate' flags for regional temperature in two regions NEU and CEU. The UKESM1-0-LL model has a single 'Inadequate' (red) flag for temperature DJF (NEU). Fig. 1 shows these temperature errors in both models to be relatively large compared to the multi-model ensemble mean RMSE. In addition, the UKESM1-0-LL model has a relatively large DJF temperature error for CEU, indicating that this temperature bias extends over two of the European land regions. These errors that are limited to specific regions may be considered acceptable for some applications, so may not necessarily always be a reason to exclude a model from a sub-selection. Here we filter the sub-set further by removing these models. Referring to Fig. 1, we can confirm that our excluded models include only those with a relatively large RMSE (1.5 times the ensemble mean) in at least one of the criteria. Also, that the eliminated models on average across the criteria have a relative error at least equal to or larger than the ensemble mean (Fig. 1). Therefore, it is again the case that the qualitative assessment has removed the models with objectively larger errors in the key criteria.

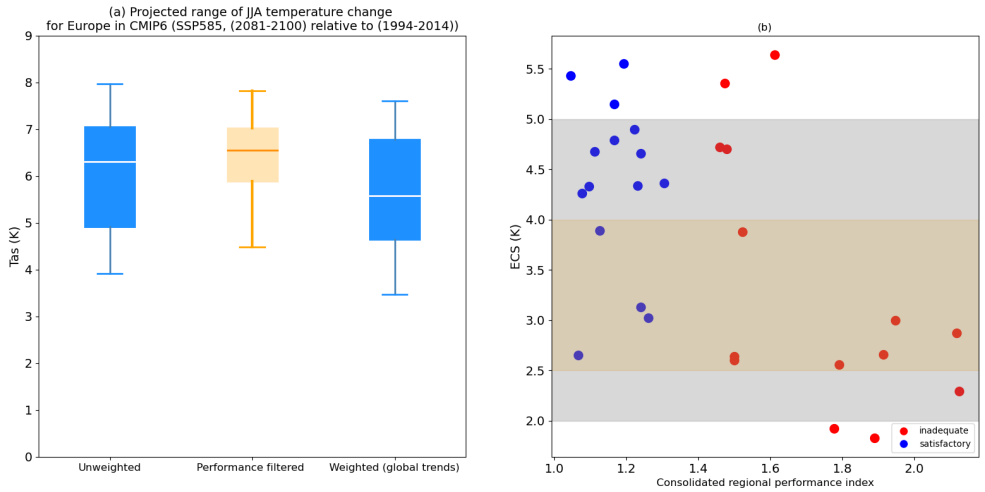


Figure 7. a) Projected range of JJA temperature change for Europe in CMIP6 (SSP585, (2081-2100) relative to (1994-2014)) for the raw unweighted multi-model ensemble, the performance filtered subset and the raw ensemble weighted for performance against global trends using the climWIP method. Boxes show 25th to 75th percentile. Whiskers show the 5th and 95th percentile. b) CMIP6 model ECS compared to consolidated regional performance index. The yellow bounds show the IPCC AR6 likely range for ECS, the grey bounds show the very likely range.

Figure 7a shows the difference in the range of projected temperature in the large-scale and regional performance filtered sub-set ~~(listed in table ??)~~ compared to the raw unweighted ensemble for JJA. The shift in the projected range for DJF in comparison is small (Fig. ~~SM2~~S2). The lower part of the range is substantially reduced for the process performance filtered in JJA (Fig 7a).

This emergent relationship between the retained models ~~tending to be~~ containing more of those with higher sensitivity contrasts with the existing literature of observational constraints on regional climate projections in CMIP6. There is an existing literature that has used ability of CMIP6 to capture either regional temperature trends ~~(e.g. ?)~~ (e.g. Ribes et al., 2022) or global trends (Liang et al., 2020; Ribes et al., 2021; Tokarska et al., 2020; Brunner et al., 2020b) that down-weight models with larger climate sensitivities, in favour of models with more modest climate sensitivities. We illustrate the contrast between this existing literature and our ~~emergent relationship~~results, by using the methodology of Brunner et al. (2020b) to illustrate the typical constraint on projections from this literature.

We use the method of Brunner et al. (2020b) (see section 3.3), applied for the global temperature trend to calculate performance weights for each model using the first ensemble member. These weightings shift the the projected temperature range downwards compared to the unweighted raw ensemble (Fig. 7 a). Our emergent relationship between less robust regional projections and lower sensitivity models was unexpected ~~,~~ and represents an apparent tension with the existing observational constraint literature based on temperature trends.

A regional consolidated performance index was created by giving the ~~satisfactory~~ 'Satisfactory' (white), ~~unsatisfactory~~ 'Unsatisfactory' (orange) and ~~inadequate~~ 'Inadequate' (red) flags a numerical score of 1 for 'Satisfactory', 2 for 'Unsatisfactory' and 3 for 'Inadequate'. The overall score for each model was then averaged by the total number of assessed criteria, to give an indication of how the model performed overall. Many of the models that performed well for the process based criteria do not fall within the IPCC AR6 likely range for equilibrium climate sensitivity (ECS) (Forster et al., 2021) (Fig. 7b).

Our result does not include any consideration of climate sensitivity and while these models are identified here as performing relatively well in a process-based assessment, the sub-set temperature range shown in Fig. 7 should not be viewed as a constraint that gives a more accurate projected range for Europe. Here we only highlight that more of the models that perform well in terms of regional physical processes have a higher climate sensitivity. It may be appropriate to select only the better performing the models from within the very likely IPCC range for ECS, or to retain just one of the models above this range to account for a higher impact scenario. It may also be appropriate to select models that are 'marginal' from the lower part of the IPCC very likely range. Alternative using an approach that considers regional impacts using Global Warming Levels could be applied to the sub-set, this is discussed further in section 5.

4.3 Sub-selection for performance and model diversity

In this section we consider how a sub-selection of a small number of example models that represent the broader characteristics of the wider ~~(satisfactory)~~ filtered projection spread could be carried out~~(point 4a) in section 1)~~. In this example application we look for a sub-set of GCMs that are both ~~plausible~~ in our filtered sub-set and sample this spread. The motivating criteria is to identify models that perform well across the whole European domain and retain as much of the spread of future projections as possible. Such an approach might be adopted by those looking for a smaller subset to drive downstream models. For example, as a selection tool for a potential Regional Climate Model (RCM) matrix, as data for a pan-European assessment of food security, or any other impact needing pan-European physically coherent climate projections, where the GCMs then would provide the climate driving data.

The models from the process performance ~~subset (table ??)~~ sub-set are placed into clusters of models that had clear dependencies (table 1). The Euclidean distance of the models is determined using the ClimWIP method (see section 3.3), for the comparison of model independence (Brunner et al., 2020b). Fig.~~SM3~~ S3 shows the independence matrix for the different models, which was used to create clusters of models that had dependencies, models with a Euclidean distance of =<0.6 were combined into clusters. Three models were found to not have a sufficient dependency to the other models to be placed in any cluster (see table 1). In most cases many of the models with similarities are from the same institution or are known to share significant code components, such as the same atmosphere model in the ~~HadGEM-GC31~~ HadGEM-GC-3.1 models and ACCESS-CM2.

In this application, to maintain model diversity as far as possible, one model was selected from each of the clusters (and two from models that fell into no cluster). Using Fig. ~~SM4~~ S4 to determine where the models are situated in the projected temperature and precipitation range for each region, these individual models are also selected to include as much of the temperature and precipitation range of the filtered multi-model ensemble as possible. The selection chosen for this example is illustrative and it may be appropriate to sub-select differently depending on the application of the sub-selection. The selected models for this example are shown in blue (Fig. 8).

In this section we have shown one example of sub-selection of a smaller ~~subset~~sub-set, using the filtered models from the previous section. There are a number of different smaller ~~subsets~~ sub-sets that could be selected using the information from the assessment tables (Figs. 5~~and ??~~). Depending on the application of the sub-selection a different approach, for example, one that includes plausible outliers (e.g., models that do not have red flags in large scale criteria), may be more appropriate in order to sample high impact, low probability regional responses.

Table 1. Table showing models clustered based on ~~euclidean~~ Euclidean distance. * models were not found to have sufficient dependencies to be placed in a cluster. Selected models are shown in bold.

No cluster*	cluster 1	cluster 2	cluster 3	
~~BCC-CSM2-MR~~ **BCC-CSM2-MR**	~~GFDL-ESM4~~ **GFDL-ESM4**	~~EC-Earth3~~ **EC-Earth3**	CESM2	
~~MRI-ESM2-0~~ **MRI-ESM2-0**	GFDL-CM4	EC-Earth-Veg	~~CESM2-WACCM~~ **CESM2-WACCM**	
MPI-ESM2-HR			TaiESM1	~~CNRM-CM~~

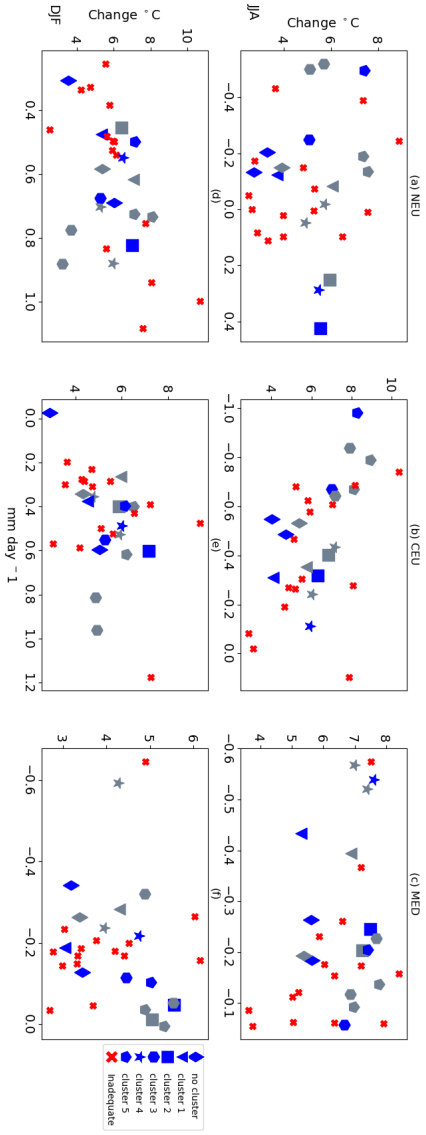


Figure 8. Temperature and precipitation projection range (SSP585, 2018-2100 relative to 1995-2014) for CMIP6 multi-model ensemble. Excluded models are shown as red. Models selected from each of the 7 clusters in table 1 shown as blue. Models from the process performance filtered subset not selected shown in grey Models from the same cluster are indicated by symbol.

5 Discussion

An overall aim of this study is to provide an assessment of CMIP6 models that can be applied by users that wish to create filtered sub-sets for Europe, for a range of applications and also wish to remove the least representative models. The assessment information could be applied to a filtering approach that is tailored depending on the criteria of interest. The assessment used in this study ~~contains both subjective and objective assessment criteria~~combines a qualitative and quantitative approach. To some extent there is always a degree of subjectivity when grading models for performance, even where more objective techniques are used, such as clustering based on evaluation statistics, (e.g. seasonal RMSE, correlation, bias as used for the blocking frequency here) there is still the difficulty of identifying where the thresholds are for what is a ~~'Satisfactory' or 'Inadequate'~~ 'Satisfactory' or 'Inadequate' model should lie, and the assessment of the relative importance of one metric versus another (Knutti, 2010). The assessment of a ~~satisfactory~~ 'Satisfactory' model will also inevitably be relative to the performance of other models in the ensemble~~to some extent~~. In the approach shown here, where the quantitative thresholds were used to guide the model classifications, these thresholds were ~~initially~~ largely determined from the distribution of performance for the ensemble. ~~If the same performance assessment that has been used here, were applied to the CMIP5 ensemble, it is likely that a much larger portion of the models would be excluded. There is a clear indication of improvements in the regional processes for CMIP6, reflecting the improvements as documented in the direct comparisons of Borchert et al. (2021b), Bock et al. (2020) , Davini and d'Andrea (2020), Priestley et al. (2020).~~ It has also been an aim throughout to maintain consistency in the way that the classifications are applied for each of the assessment criteria. In practice this can be difficult to achieve as many of the GCMs generally capture some of the large-scale processes relatively poorly (e.g. blocking frequency and CEU precipitation) in comparison to others for each criteria, and also due to the difficulty in evaluating others (e.g. AMOC).

A further challenge is that not all models have been assessed against all criteria. Analyses that assessed storm tracks, blocking frequency and the AMOC provide valuable further information regarding the performance of the models, but were not available for every model in the study, therefore it was necessary to consider whether a model should be eliminated on the basis of one of these criteria, when other models where their performance was unknown may be kept in the selection. It was found to be the case that the flags for exclusion did not occur in isolation, severe errors (red flags) for large scale circulation, storm tracks and blocking frequency often occurred in more than one ~~criteria~~ criterion (or in some cases alongside multiple orange flags). Severe errors (or ~~Inadequate~~'Inadequate', i.e., those flagged red) in the AMOC, another ~~criteria~~ criterion where data was limited, ~~also did not occur in isolation and were found to occur in a few models that had also been flagged as Inadequate (red flagged) due to poor circulation features and in particular the poor representation of storm tracks. For example~~were due to a very weak representation of this feature and where this is the case a severe cold bias in the SPG region was also present (NESM3).

Considering the regional impact of eliminating the models flagged as 'Inadequate' (flagged red, Fig. 5); the lowest temperature response models are excluded in summer (JJA) for the NEU, CEU and MED (Fig. 8). For summer rainfall, in the ~~case of a~~

~~red flag for a very weak AMOC in the NESM3 model, severe winter cold bias (red flagged) in the NA SST were also found to be present~~NEU and MED, many of the models showing a more neutral change in rainfall are excluded. Greater warming is generally linked to stronger summer drying and increased winter precipitation. The exclusion of many models with a more modest projected temperature increase also excludes many of the more neutral projected changes in precipitation.

Filtering from the CMIP6 ensemble by excluding the least realistic models for Europe leads to the removal models throughout the projected temperature range ~~,~~ but removes more of the models that have a more modest response. The retention of higher sensitivity models is ~~an emergent consequence of assessment of skill at~~ due to more of the higher sensitivity models demonstrating a greater skill for reproducing regional processes. The revised temperature projections for the filtered GCMs for each region leads to a shift upwards in the median of the projected JJA temperature range, due to more of the higher sensitivity models performing well against the ~~process based~~ process-based criteria (Fig. 7). This ~~represents~~ may represent a particular challenge for potential applications ~~wanting to sample~~ where sampling regional climate responses in the lower end of the IPCC climate sensitivity ~~range. Many~~ (ECS) range is required, as many of the CMIP6 models in the lower part of the ECS likely range were excluded by our processed based assessment (Fig. 7b).

~~We see this same finding feed through to our example of performance filtered models, where the lowest temperature responses models are excluded for summer in all European regions (figure 7). Considering the regional impact of eliminating the models flagged as Inadequate (flagged red, Fig. 5); the lowest temperature response models are excluded in summer (JJA) for the NEU, CEU and MED (Fig 8) . For summer rainfall, the impact of performance filtering is less clear in the NEU and MED, however many of the modelsshowing a more neutral change in rainfall are excluded. Greater warming is generally linked to stronger summer drying and increased winter precipitation. The exclusion of many models with a more modest projected temperature increase also excludes many of the more neutral projected changes in precipitation~~Using the IPCC AR6 likely range for ECS (and or TCR, Hausfather et al. (2022) has also been suggested as an approach to model screening for the CMIP6 ensemble. Other regional sub-selection studies for CMIP6 have eliminated models with high global sensitivity (Mahony et al., 2022). Any assessment that excludes models based on both the performance criteria here and metrics like global climate sensitivity or global trend criteria (which tends to exclude models with higher climate sensitivities) will be left with only a small sub-set of "adequate" models. This apparent tension is likely to be less evident where Global Warming Levels are instead adopted. Using this approach, when adopting a selection based on climate performance, such as presented here, would enable a broader set of "adequate" realisations to be explored.

Our results contrast with the existing literature based on evaluation against historical temperature trends (Liang et al., 2020; Ribes et al., 2021; Tokarska et al., 2020). Many of the models that score well against ~~process based~~ process-based criteria, have a higher ~~climate sensitivity (ECS)~~ECS. ECS is not considered in this study as a sub-selection criterion, because the focus of this work was on the assessment of how well models capture the main regional climate processes. Links between plausibility of CMIP6 projections, based either on their historical global or regional temperature trends or climate sensitivity (Hausfather et al., 2022) are well established in the literature for CMIP6 ~~(Ribes et al., 2021; Liang et al., 2020; Tokarska et al., 2020; Sherwood et al., 2020)~~

(Ribes et al., 2021; Liang et al., 2020; Tokarska et al., 2020). When the raw ensemble is weighted against performance for global trends (Fig.7a) the effect is to shift the temperature range downwards. This shift for our raw ensemble is not as large as typically seen for in other studies for global trends (e.g. Liang et al., 2020; Ribes et al., 2021; Tokarska et al., 2020), this may be due to our using a single ensemble member for this study, some differences in methodology, or due to summer warming in Europe being thought to be about 30% higher than the annual mean global warming ~~(?). ?~~ (Ribes et al., 2022).

Ribes et al. (2022) find a constrained regional projection range for mainland France for ssp585 (5.2 to 8.2 °), that is similar to the projected range for summer of our performance filtered subset, using a combination of modelling results and observations. Our upper 95th percentile and lower 5th percentile is a little lower than this (for a pan-European range), our median for the performance filtered range is very similar to their central estimate of 6.7 °C (Fig. 7a).

~~Using the IPCC AR6 likely range for ECS (and or TCR, Hausfather et al. (2022)) has also been suggested as an approach to model screening for the CMIP6 ensemble. Other regional sub-selection studies for CMIP6 have eliminated models with high global sensitivity (Shiogama et al., 2021). Our results also highlight the tension for regional filtering of the CMIP6 multi-model ensemble between including lower sensitivity models or models within the ECS likely range, while also taking account of regional model performance for key processes (Fig 7b). Only a small number of models from our multi-model ensemble, that perform well against the regional criteria are within the ECS likely range. Due to many of lower sensitivity models (and those within the IPCC AR6 ECS likely range)struggling to reproduce the European climate~~For assessments of model performance against historical temperature trends, where the regional trends are also taken into account there may be less of a tension with our assessment, than is the case with those that are based on global trends alone (e.g. Liang et al. (2020); Ribes et al. (2021)).

6 Conclusions

We provide an assessment of regional processes and biases (~~Figs. 5and ??) ,~~ Fig. 5) for a multi-model ensemble for CMIP6 that can be used to inform sub-selection for the European region. This can be used to aid the creation of bespoke sub-selections for a particular application (e.g., sub-selection of a small number of representative ensemble members, for downscaling or impact assessments), alternatively the sub-sets that have been demonstrated here can be also be used directly.

~~The filtering~~ Filtering an ensemble of CMIP6 models based on performance against key processed based criteria results in the projected temperature range being shifted upwards. This is due to the removal of a larger proportion of the lower climate sensitivity models, that do not perform adequately against the assessment criteria. We also find that many of the higher sensitivity models score well against the ~~process based~~ process-based assessment and that these models are better able to represent the features of the European climate. It is not clear whether the emergent relationships we found (between better models and higher sensitivities) is circumstantial or reflects an underlying physical basis. If it reflects an underlying physical relationship (where atmospheric processes needed to capture regional feedbacks also drive stronger climate feedbacks) then this might

imply greater confidence in higher end regional changes. If on the other hand the sampling of higher sensitivity models is circumstantial (simply due to chance), this represents a challenge as there are few CMIP6 models that sample the central and lower end of the IPCC AR6 likely climate sensitivity range. This remains an open question, which we have not been able to resolve in this work.

Our results highlight a tension for regional sub-selection between performance against the global temperature trend and the ability of the models to capture the features of the regional climate in the CMIP6 multi-model ensemble. For cases where changes in temperature are not the only variable of interest (or the primary concern) many of the higher sensitivity models are likely to provide more reliable information regarding the future climate. Potential users of regional climate projections should be aware that there is a potential tension between constraints from large scale temperature change/climate sensitivity and the assessment of regional processes, for Europe at least.

Code and data availability. The code used to apply the climWIP method is publicly available via the ESMValTool (https://docs.esmvaltool.org/en/latest/recipes/recipe_climwip.html)

The data used in this study is available through the ESGF data portal at https://esgf-node.llnl.gov/projects/esgf-llnl/. Details of the models are available from https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_source_id.html

A version of the assessment figures used in this paper is available on github https://github.com/tepmo42/cmip6_european_assessment as a full spreadsheet of all available assessments (for Europe) carried out for CMIP6 models to date.

~~The code used to apply the climWIP method is publicly available via the ESMValTool ()~~

~~The data used in this study is available through the ESGF data portal at https://esgf-node.llnl.gov/projects/esgf-llnl/. Details of the models are available from https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_source_id.html~~

~~A version of the assessment figures used in this paper is available on github https://github.com/tepmo42/cmip6_european_assessment as a full spreadsheet of all available assessments (for Europe) carried out for CMIP6 models to date.~~

Appendix A: Appendix

A1 Annual precipitation cycle

The annual precipitation cycle was assessed as a regional criteria for each of the European land regions (Gutiérrez et al., 2021) (see Fig.~~SM1~~S1). The precipitation cycle for each model was assessed against the EOBs data as ~~monthy~~ monthly means (see figure A1) using the baseline period of 1995-2014. A combination of the correlation and RMSE for each of the 4 seasons is used to assess whether models should be categorised either as satisfactory or unsatisfactory.

In order to sort the models into categories the seasonal RMSE and correlation were used as a guide (Fig. A1 b-e), it was observed that in most regions a poor correlation with the observed cycle had a large seasonal RMSE compared to other models

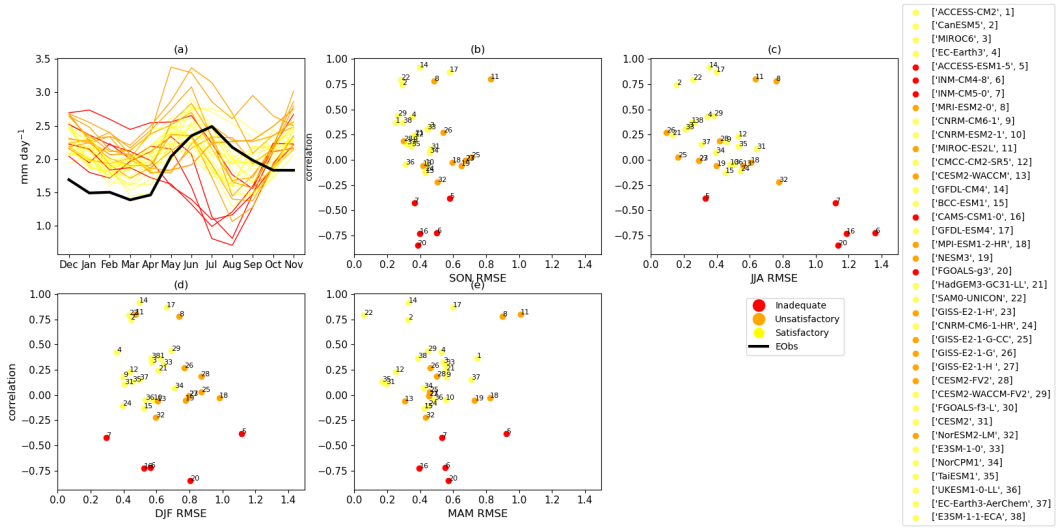


Figure A1. Precipitation annual cycle for CEU (top left), model comparison with EOBS (shown as solid black line). Correlation (over 12 months) and seasonal RMSE for each model. Monthly averages are taken over a ~~20 year~~ 20-year climatology (1995-2014). The RMSE and correlation are calculated from the monthly averages

(>0.75mm day^{-1}), in at least one season. The exception was in CEU where most models had a poor correlation with the observations. The models with a relatively low error of less than 0.6mm day^{-1} for all four seasons were classified as satisfactory. Models with a RMSE of greater than 0.75mm day^{-1} in one season were classified as unsatisfactory.

In CEU the models with a substantially negative correlation (approx -0.25) had a large seasonal RSME (>1mm day^{-1}) and very poor agreement with the seasonal cycle where they essentially showed a strong seasonal drying in the wet season for CEU (figure A1). These models are classified as excluded.

There were some models (RMSE >0.6mm day^{-1} <0.74mm day^{-1}) that did not fall immediately any classification, ~~these models were flagged for classification based on visual inspection of the annual precipitation cycle (A1a)~~Where the RMSE error in all seasons was < <0.74mm day^{-1}, the model performance was generally satisfactory. However, in the case of some models closer this threshold with some were 'Unsatisfactory', if they had a lower correlation with the EOBS data.

A2 Sea Surface Temperature bias

Seasonal average Sea Surface Temperatures (SSTs) were assessed for each of the models using the HadISST1 reanalysis (Fig A2) for the baseline period 1995-2014. Surface skin temperatures from the atmospheric models were used, the corresponding ice concentration fields from the atmosphere model were only available for a smaller number of models. To estimate the ice extent and avoid errors in the ~~calculation~~ assessment of the SST bias in areas affected by ice, a seasonal average ts < 0 is used as a proxy for the 5% ice concentration to mask these areas. The areas masked by this proxy is compared to the extent of the 5% ice concentration in the models and found to be a good approximation. As the area affected by ice is approximated this is not compared directly to the 5% ice field from HadISST1 for the ~~asssessment~~assessment, however where the mask areas are significantly larger than the 5% ice concentration in HadISST1 (Fig A2, bottom right), a large cold bias in these areas is inferred (figure A2). This bias in sea ice and the SST surrounding northern Europe is found to be well captured by the large-scale near surface temperature bias (see section A3). Therefore it is noted here, as an important consideration for the European climate, but not included explicitly in the assessment of the NA SST error classifications. For the NA SST assessment we focus on errors in key regions of the NA for the European climate.

The NA SST assessment is based on two key areas of the NA, the subpolar gyre (SPG) and Gulf Stream northwest corner (GS) regions. These have been selected from Ossó et al. (2020), who identified a North West region of the North Atlantic GS as important for weather patterns over Europe, and Borchert et al. (2021b) to define the SPG region, which has previously been shown to modulate the probability of occurrence for summer temperature extremes in central Europe (Borchert et al. (2019) ; Fig. S7). These regions as well as their gradient has been demonstrated to carry relevance for dynamical atmospheric influences of NA SST on European summer climate (Carvalho-Oliveira et al., 2022)), highlighting their relevance in the context of this study. During a qualitative inspection of the models (see A2) these regions were also identified as often areas to routinely show a substantial bias in the models.

A small number of models had extensive areas with a very large winter negative SST bias (Fig A2, bottom row), this results in a substantial over estimation of winter ice extent to the north of Scandinavia and around Greenland. ~~E3SM-1-ECA~~ NESM3 and CAMS-CSM1-0 have a large widespread negative bias that extends beyond the the regions of sea ice to the NA and SPG. ~~These are examples of excluded models~~ (Fig A2). The models with the largest SPG RMSE are NESM3, CanESM5, CAMS-CSM1-0 (shown in Fig 1) and FGOALS-f3-L. In addition, FGOALS-f3-L has an RMSE for the GS region of more than twice the ensemble mean RMSE. These models are all flagged as 'inadequate'.

~~For DJF a~~ A number of models also had areas with substantial but limited areas of warm bias (>6K) in the area around the Gulf Stream and larger areas in the SPG (>3K) e.g. CESM2, ~~NorCPM1~~INM-CM50, NorESM2-LM (Fig A2). In addition, these models also have areas cold bias in the SPG, this combination of warm and cold biases in different areas also results in a poor representation of the SPG temperature gradient. These ~~errors are not considered large enough to exclude these models~~

but they have been flagged as unsatisfactorymodels also had an RMSE larger than the ensemble mean in the SPG assessment region and were classified as 'unsatisfactory'. The INM models are an exception in the 'unsatisfactory' category in terms of not having a large SPG error, however these two models have some of the largest errors in the GS region, with the exception of FGOALS-f3-L.

Satisfactory models had a lower bias in all areas, some had small areas with a larger bias (often around the Gulf Stream or in some parts of the SPG, (e.gAWI-CM-1-1MR,. ACCESS-CM2), but the affect effect of these areas did not prevent a reasonable representation of the SPG gradient. The models classified as 'satisfactory' all have a RMSE in the assessed region that is less than the ensemble mean.

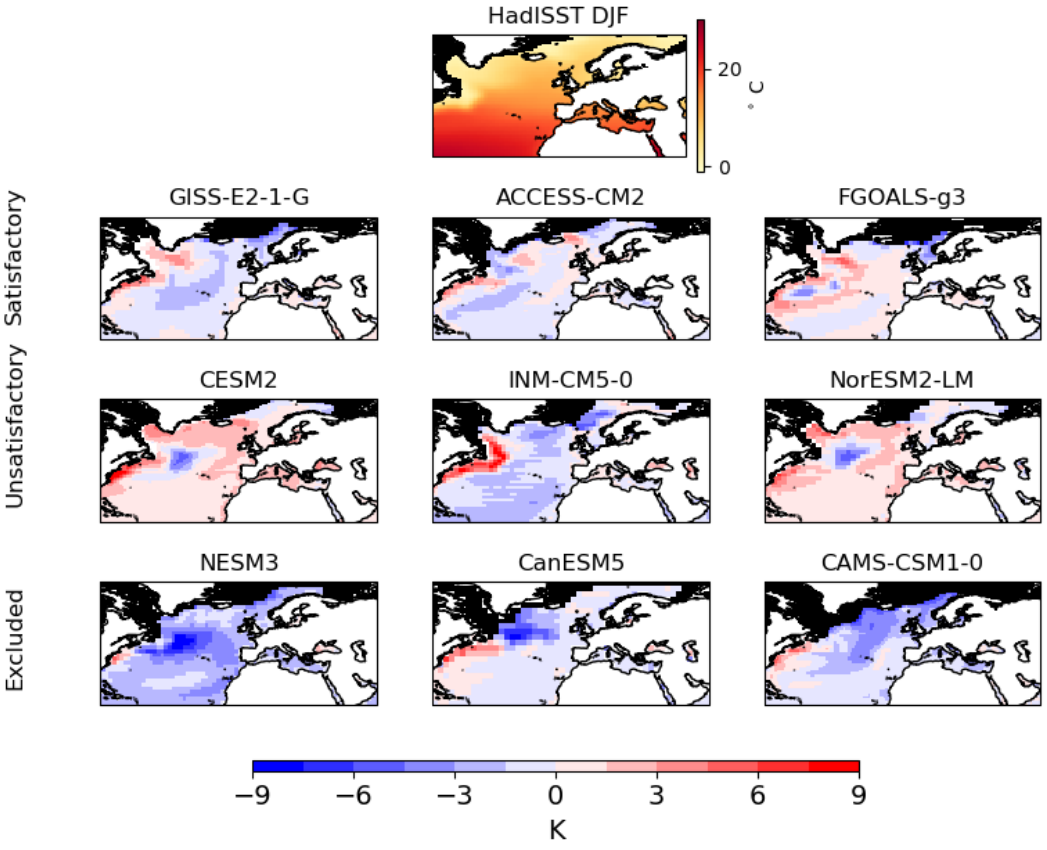


Figure A2. Model SST bias (compared to HadISST). Seasonal average calculated for a 20 year climatology (1995-2014). Areas where the model SST < 0 are mask in black (this was found to approximate to 5% ice concentration). Top row shows the HadISST and 5% ice concentration field.

In JJA the satisfactory models again had smaller areas of bias (>3K) around coastal regions, but these were not widespread (Fig. A3). ~~The unsatisfactory models had larger regions with a substantial cold bias , often with regions <0°C, that indicate over-estimation of the summer ice extent (e. g.~~ Models flagged as satisfactory also had SST RMSE in the SPG and GS regions that were less than the ensemble mean error. Although the model errors in the SPG ans GS are satisfactory there is a cold bias in the SSTs in the Norwegian and Barents Sea in the FGOALS-g3 (and GISS-E2-1-G~~, NorCPM1). The models that were excluded show a more severe cold bias , with in some cases sea-ice estimated near European land regions (e. g.~~) model. In the FGOALS-g3 ~~).~~ model there is also an excess in the sea ice extent in the Barents region. While the SST assessment has been focused on the NA SST region it is noted that biases in this region are also important for the European climate. These biases are captured in the model classification for temperature bias (which includes near surface temperature bias over the ocean).

The unsatisfactory models had larger regions with a substantial cold bias in the SPG and/or larger biases in the GS region, that were larger than the ensemble mean. The CAMS-CSM1-0, CanESM5 and FGOALS-f3-L models with the largest SPG errors were classed as 'inadequate'.

A3 Near surface temperature bias

A3.1 large scale bias

The model near surface temperatures ~~are~~ is compared to ERA5 reanalysis (Fig A4) for the baseline period 1995-2014. These were assessed ~~as a~~ on the large scale (including surrounding areas over the NA, Norwegian Sea, Barents Sea and nearby Arctic regions) criteria (Fig.A4,) and also more specifically for the land points of each SREX region ~~. The latter was included as a regional flag for DJF and JJA, these are discussed in the next section~~ (see in the following section A3.2).

For the large scale assessment there is inevitably some overlap with the assessment of SST temperatures as near surface temperature over North Atlantic regions is taken into account. The large scale qualitative assessment considers whether there is widespread areas of temperature bias in land regions of Europe or in other regions where they could be expected to have downstream impacts, e.g. nearby land areas ~~or NA~~NA, or other ocean regions nearby the European land areas. A more widespread bias as opposed to a smaller more regionally based temperature bias, indicates an issue with the ~~large scale~~ large-scale processes that will affect all the European regions, while a more local area of bias is likely to indicate issues related to processes in a particular region. Where biases in land regions are found in more than one European region however these are likely to also indicate an issue that may affect the whole European area. The RMSE error over the region as a whole and each of the land regions is taken into consideration for the model calculation alone with a more qualitative assessment of regions of bias.

For JJA, MIROC6 has a large widespread positive summer bias over European land regions, north Africa and Greenland, this bias is largest in the CEU and MED, but this is not extended over the NA where there is a cool bias. The warm bias in the MED and CEU regions is exceptionally large (>8K in some areas), but it is not limited to these regions, with a smaller but

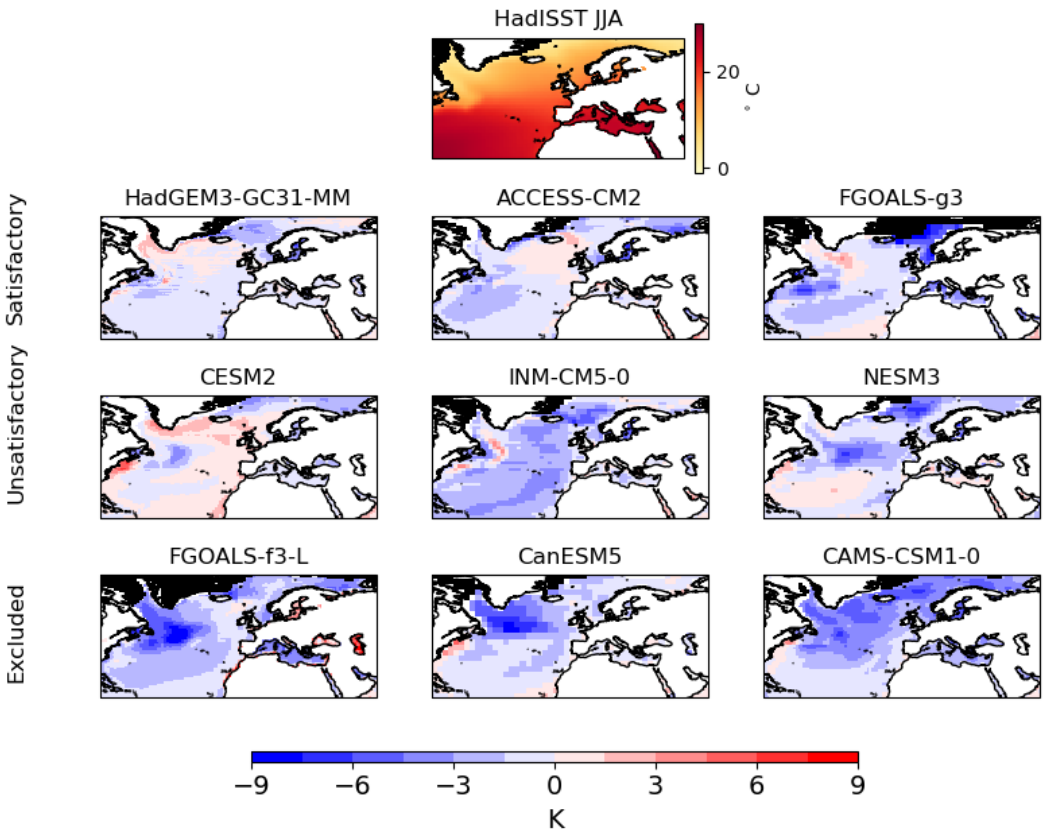


Figure A3. Model SST bias (compared to HadISST). Seasonal average calculated for a ~~20 year~~ 20-year climatology (1995-2014). Areas where the model SST < 0 are mask in black (this was found to approximate to 5% ice concentration). Top row ~~,~~shows the HadISST and 5% ice concentration field.

still substantial bias for all land regions (Fig A4). The RMSE is the largest in the model ensemble for the whole region and the northern and central regions. MIROC-ES2L has a similar pattern of errors as MIROC6 (although not quite as large, still more the 1.5 times the ensemble mean RMSE. CAMS-CSM1-0 has a large widespread bias negative bias in all areas of Europe, this model also has a large cold bias in ~~DJF~~ JJA for both for land and SST. It has one of the largest RMSE for the large-scale region and the largest in the ensemble for northern Europe. FGOALS-g3 also had widespread biases with an unusual pattern showing an area of exceptionally large cold bias to the north of Scandinavia and the UK (>8K), while also having a substantial warm bias in the eastern area of CEU (4-6K around the black sea area). ~~These models were excluded based on these large errors in seasonal meantemperature.~~ The RMSE for the whole region is above average, but not exceptionally large compared to the rest

of the ensemble, this is largely due to a relatively small bias in the NA, as noted in the SST assessment. The RMSE error in the central European region is more than 1.5 times the ensemble mean. The additional area of large low bias in the Norwegian and Barents Sea area, with the resulting excessive sea ice (see Fig.A3), has led to this model also being rated as 'inadequate'. The INM-CM4-8 model has a large positive bias in both the central and Mediterranean regions, and RMSE for both these regions and the SPG is more than 1.5 times the ensemble mean error; therefore this model has also been classified as 'inadequate'.

Examples of models classified as ~~unsatisfactory~~ 'Unsatisfactory' for JJA bias, include NESM3, GISS-E2-1-G and ~~E3SM-1-1-ECA~~ INM-CM4-8 (Fig A4). NESM3 has a substantial warm bias in eastern CEU and MED regions (> 4 in some areas) and areas of cold bias in the NA (4-7K). GISS-E2-1-G ~~and E3SM-1-1-ECA also both had~~ has substantial more widespread areas of cold bias (Fig. A4)~~The E3SM-1-1-ECA model had a large area of substantial cold~~ . The INM-CM5-0 model has a substantial warm bias in the ~~NA that extended to European land areas, although the largest areas of bias were not European land regions these regions of bias can be expected to have downstream impacts~~central European region and SPG area. Its overall RMSE for the large-scale area is large than the ensemble mean RSME, this model is classified as 'unsatisfactory'.

Examples of 'satisfactory' models with a bias of </= 2K in most regions for JJA and limited regions with bias of up to 4K in limited areas, include GFDL-CM4, CNRM-CM6-1-HR and EC-Earth3 (Fig. A4 (top row)) Models classified as 'satisfactory' had a large scale RMSE that was less than or close too (slightly above) the ensemble mean RMSE.

For DJF the cold bias in the ~~excluded models~~ models that are classified as inadequate is pronounced, especially in northern European areas (Fig. A5). These models all had an RMSE for the large-scale area that was more than 1.5 times the ensemble mean RMSE. In the case of FGOALS-g3 it was more than twice the ensemble mean error.

The unsatisfactory models included those with substantial cold bias in areas than while not directly over European land regions can be expected to have some downstream impacts on them (e.g., NESM3, ~~GISS-E2-1-g~~GISS-E2-1-G). In ~~some~~ several cases substantial biases are present in the land regions of interest (e.g., NorESM2-LM). The ~~satisfactory models in some cases still show some areas of bias, particularly cold bias over Scandinavia (e. g. KACE-1-0-G) or warm bias in Eastern CEU (e.g. MRI-ESM2-0) . However these bias's were common in the GCMs. Some of these models with local areas of bias that were found to be satisfactory in the large scale assessment were flagged in the following European land regions assessment~~models classified as 'unsatisfactory' all had RMSE errors large than the multi-model mean. The only exception is the UKESM1-0-LL, which had RMSE for the large-scale area that was slightly lower, but substantial errors in two European land regions (northern and central Europe) that were among the largest in the multi-model ensemble. 'Satisfactory' models had smaller biases over all regions and a RMSE for the large-scale that was smaller than the multi-model ensemble mean.

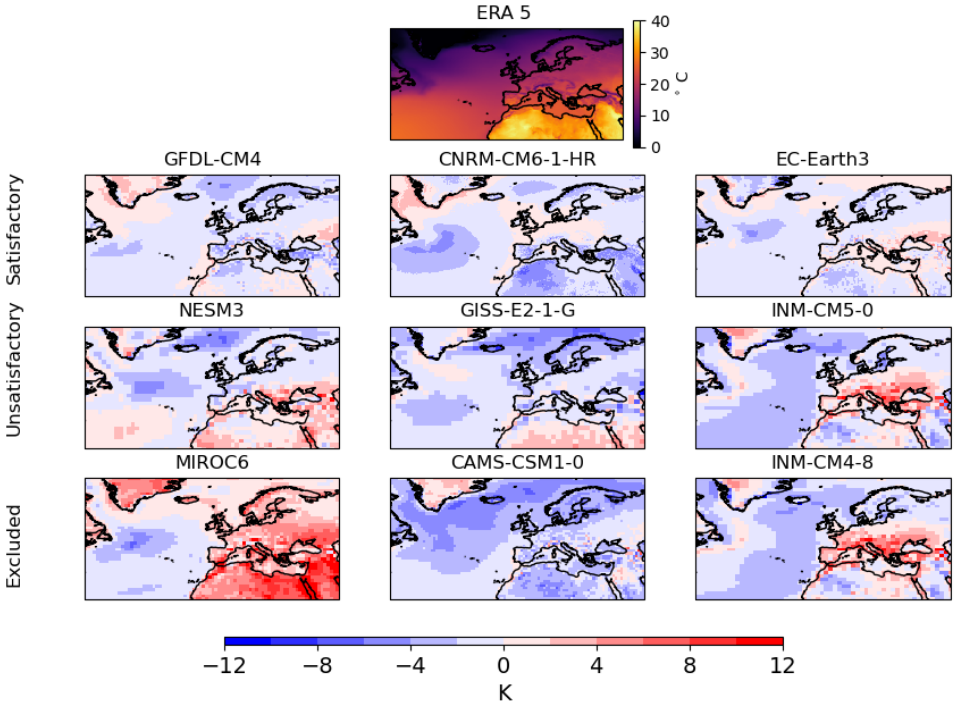


Figure A4. Model large-scale temperature bias. Seasonal JJA average calculated for a 20 year climatology (1995-2014).

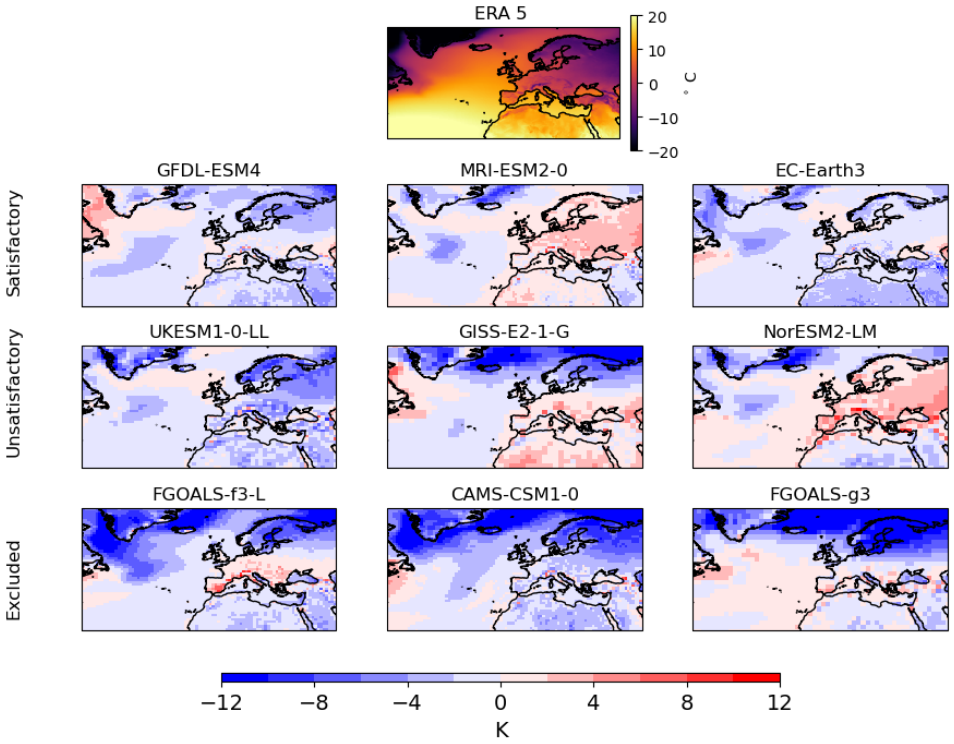


Figure A5. Model large-scale temperature bias. Seasonal DJF average calculated for a 20 year climatology (1995-2014).

A3.2 European land regions

In addition to the large scale assessment the three IPCC AR6 land regions (Gutiérrez et al., 2021) were individually assessed to identify land areas of seasonal temperature bias. The spatial mean seasonal RMSE for all land points in each region was calculated and used as a guide for assessment, along with a visual inspection of the spatial temperature bias. A small number of models were classified as ~~excluded~~ 'Inadequate' for individual regions due to areas with a large local bias, that were not excluded due to a temperature bias in the ~~large scale~~ large-scale assessment. These models ~~were classified as excluded~~ may be considered as 'Inadequate' only for the region.

~~For JJA, MIROC-ES2L and INM-CM4-8 have a regional RMSE~~ The RMSE for each region is used to classify the models, for JJA the thresholds were for 'satisfactory' < 2.5K, 'unsatisfactory' >2.5K but < 4K, 'inadequate' >4K~~with several areas showing a large warm bias. The INM-CM5-0 model has an average RMSE~~ .As is the case in determining any threshold there is a degree of subjectivity and these thresholds are based on the relative performance of the models across the ensemble. For DJF the thresholds were the same except that the threshold for 'inadequate was increased to >~~4k for both CEU and MED. It is noted that INM-CM4-8 and INM-CM5-0 models were also unable to represent the precipitation annual cycle in the CEU region, with the JJA season too dry, this may be related to the warm seasonal bias in the CEU and MED. These MIROC-ES2L , INM-CM4-8 and INM-CM5-0 were flagged as excluded for MED due to summer temperature bias and INM-CM5-0 was also flagged as excluded for CEU.~~5K.

~~UKESM1-0-LL and KACE-1-0-G had a large cold DJF bias (>8K in some coastal grid points) in northern parts of Scandinavia, with a RMSE of 5K-5.5K, these models were classified as unsatisfactory for NEU.~~

A4 Atlantic Meridional Overturning Circulation

The representation of the AMOC is still considered to be deficient even in state of the art GCMs, where its associated climate impacts are also thought to have been underestimated (Zhang et al., 2019). In addition due to the limited availability of observational data there is still considerable uncertainty in the recent AMOC evolution (Menary et al., 2020), and accurate assessment of the AMOC in climate models remains challenging. For this study some assessment of the AMOC is considered to be important due to its potential role in future changes in the European climate. The aim is to identify and flag the poorest models with large errors in the representation of the AMOC compared to the observational data from the rapid array. Examples of the overturning stream function for each model shown (Fig A6) is calculated using the method of (Menary et al., 2020).

NESM3 and IPSL-CM6A-LR both show poor agreement with the observational data, with a consistently weak AMOC (Fig A6)~~, these models are classified as excluded.~~ . NESM3 was classified as 'Inadequate', the impact of the AMOC on the NA SSTs is also flagged due to a large cold bias. The AMOC for IPSL-CM6A-LR is flagged as 'Unsatisfactory', the error may impact on the representation of the NA, but the impact on the reliability of future projections is not clear, a similar error was present in CAMS-CSM1-0, which is also flagged as 'Unsatisfactory'. In contrast the NorESM2-LM model has a consistently

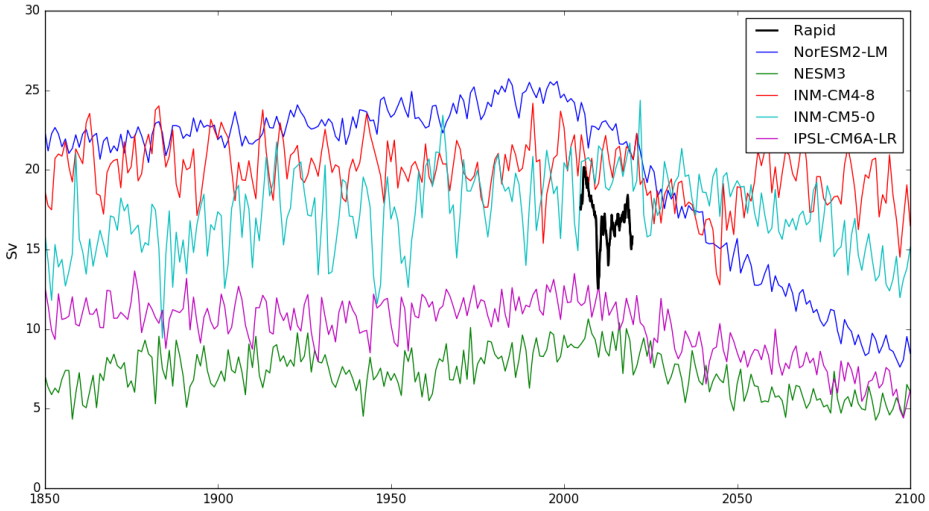


Figure A6. AMOC strength at 1000m (from v - velocities compared to rapid array (annual mean Sv) at 26°N, AMOC data from Menary et al. (2020)

strong AMOC through the historical period, with a rapid decrease in more recent years which is not seen in the observational data, this model is also classified as ~~excluded as was the GISS-E2-1-G due to a strength of the AMOC well in excess of that observed~~'Unsatisfactory'. The other models for which AMOC data was available are classified as satisfactory (e.g.INM-CM4-8 and INM-CM5-0), as they do not show a large deviation from the observations~~, the one exception is CAMS-CSM1-0 which was classified as unsatisfactory due to a substantially weak AMOC~~.

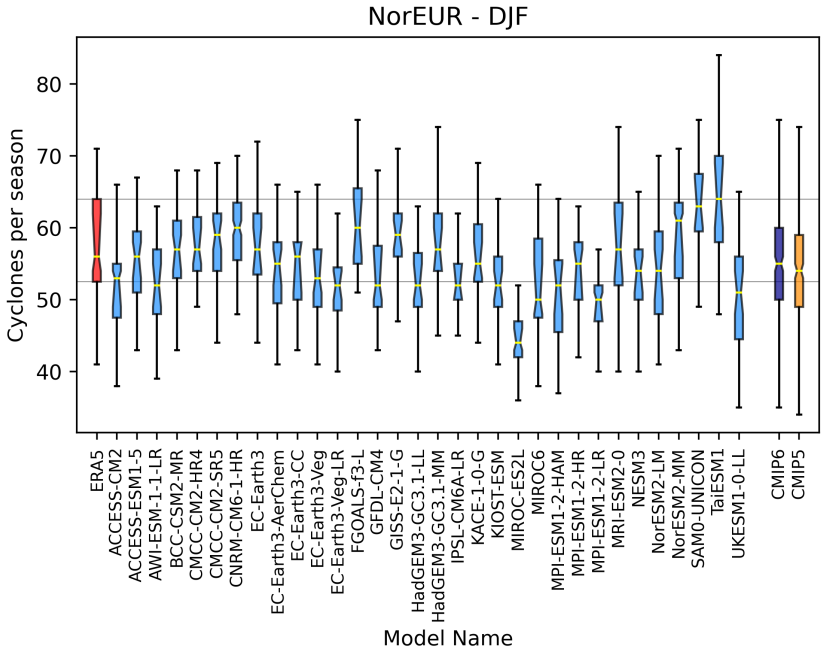


Figure A7. 'Boxplots of cyclone numbers per DJF season for NEU region (co-ordinates). Boxes are shown for ERA5 (red), 32 CMIP6 models (light blue), and the CMIP6 (purple) and CMIP5 model ensembles (orange). Boxes extend to the 25th and 75th percentile of the distributions, with whiskers extending to 1.5 times the inter-quartile range. Horizonal yellow lines indicate the medians. Notches around the medians show its uncertainty based on 10,000 random resamples. Horizontal gray lines indicate the ERA5 25th and 75th percentiles

A5 Storm Tracks

A5.1 Regional assessment

The storm tracks were also assessed regionally to determine whether the number and variability of the cyclones in a particular region were captured satisfactorily by the models. This used the analysis of Priestley et al. (2020) for the individual European regions. The baseline time period used for this assessment is 1979/80-2013 for CMIP6 (1979/80-2004/05 for CMIP5) and the model data is compared to ERA5.

Where the 25th and 75th percentile of the range ~~overlaped~~ overlapped and was similar in size to the ERA5 data the model was classified as satisfactory. If the ~~the~~ interquartile range of the model had no overlap with ERA5 data or the size of the interquartile range was substantially smaller ~~then~~ than the model was categorised as unsatisfactory for the region (see Fig.A7. Models ~~were not excluded on the basis of regional analysis~~are not excluded based on the regional analysis, these flags are for information only.

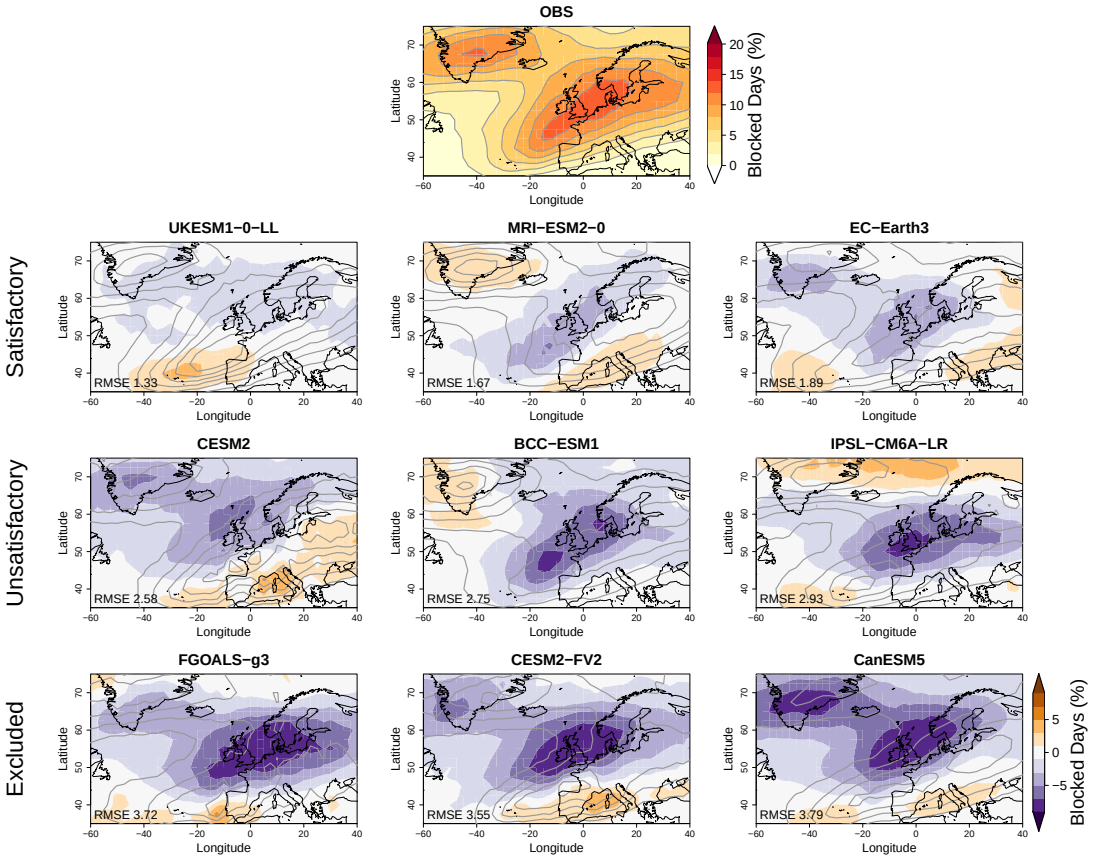


Figure A8. Examples of DJF Blocking Frequency classifications for a sample of individual models..

A6 Blocking frequency

Atmospheric blocking is a recurrent weather pattern typically occurring in the mid-latitudes at the exit of storm track (Rex, 1950; Pelly and Hoskins, 2003). It is characterized by a high-pressure low-potential vorticity quasi-stationary large-scale anomaly which is able to "block" or divert the movement of the traveling cyclones, creating anomalous weather in its underlying region. One challenging issue for the climate community is the struggle that weather and climate models have in reproducing the observed frequency of atmospheric blocking (D'Andrea, 1998; Masato et al., 2013). Indeed, state-of-the-art climate models are known for underestimating the frequency of atmospheric blocking especially over the Euro-Atlantic sector, albeit notable improvements have been observed with the last generation of models (Davini and d'Andrea, 2020).

In this work atmospheric blocking is identified with an objective index based on the reversal of the daily geopotential height gradient measured at 500hPa, making use of the blocking index developed by Davini et al. (2012)). The index is the 2-d extension from 30°N to 75°N of the canonical definition by Tibaldi and Molteni (1990). However, we here adopt a blocking

definition which includes a third supplementary condition south of the blocked region aimed at excluding the low latitude blocking events (see Davini et al. (2012) for details). Defining Z500 as the daily geopotential height at 500hPa interpolated on a common regular 2.5°x2.5°grid, three meridional gradients are considered:

$$GHGS(\lambda_0,\phi_0) = \frac{Z500(\lambda_0,\phi_0) - Z500(\lambda_0,\phi_S)}{\phi_0 - \phi_S}, \tag{A1}$$

$$GHGN(\lambda_0,\phi_0) = \frac{Z500(\lambda_0,\phi_N) - Z500(\lambda_0,\phi_0)}{\phi_N - \phi_0} \tag{A2}$$

$$GHGS2(\lambda_0,\phi_0) = \frac{Z500(\lambda_0,\phi_S) - Z500(\lambda_0,\phi_{S2})}{\phi_S - \phi_{S2}} \tag{A3}$$

and ϕ_0 ranges from 30°N to 75°N while λ_0 ranges from 0° to 360°. ϕ_S= ϕ_0 - 15°, ϕ_N= ϕ_0 + 15°, ϕ_{S2}= ϕ_0 - 30°. Instantaneous Blocking is thus identified when:

$$GHGS(\lambda_0,\phi_0) > 0 \qquad GHGN(\lambda_0,\phi_0) < -10 \text{ m}/^\circ\text{lat} \qquad GHGS2(\lambda_0,\phi_0) < -5 \text{ m}/^\circ\text{lat} \tag{A4}$$

As done by Davini and d'Andrea (2020), no spatial or temporal filtering is applied.

29 CMIP6 models are taken into consideration, considering the time window 1961-2000. In order to define an objective method to classify into categories the atmospheric blocking bias over the Euro-Atlantic region (60°W-40°E, 35°N-75°for winter and 60°W-40°E, 45°N-75°N for summer) two basic metrics has been introduced: the RMSE and Pearson correlation coefficient, evaluated against ERA5 reanalysis. Both RMSE and Pearson correlation coefficients are then standardized and used as non-dimensional parameters to perform a k-means clustering (Michelangeli et al., 1995) with k=3. In this way, climate models showing similar bias in both magnitude and pattern are clustered together, taking into account not only the size of the bias but also its shape. An example of the classification is provided in Figure A8.

Author contributions. Tamzin Palmer: conceptualization, data curation, formal analysis, Investigation, methodology, software, supervision, validation, visualization, writing - original draft, writing -review and editing.

Carol McSweeney: conceptualization, funding acquisition, methodology, project administration, supervision, validation, writing -original draft, writing - review and editing.

Ben Booth: conceptualization, funding acquisition, methodology, project administration, supervision, validation, writing -original draft, writing - review and editing.

Matthew Priestley: conceptualization, data curation, formal analysis, software, Investigation, methodology, validation, visualization, writing - original draft, writing -review and editing.

Paolo Davini: conceptualization, data curation, formal analysis, software, Investigation, methodology, validation, visualization, writing - original draft, writing -review and editing.

Lukas Brunner: conceptualization, methodology, validation, visualization, software, writing -review and editing.

Leonard Borchert: conceptualization, methodology, validation, writing -review and editing.

Matthew Menary: validation, data curation, software.

~~Tamzin Palmer: conceptulization, data curation, formal analysis, Investigation, methodology,software, supervision, validation, visualization, writing - original draft, writing -review and editing Carol McSweeney: conceptulization, funding acquisition, methodology, project administration, supervision, validation, writing -original draft, writing - review and editing. Ben Booth: conceptulization, funding acquisition, methodology, project administration, supervision, validation, writing -original draft, writing - review and editing. Matthew Priestley : conceptulization, data curation, formal analysis, software, Investigation, methodology, validation, visualization, writing - original draft, writing -review and editing Paolo Davini: conceptulization, data curation, formal analysis, software, Investigation, methodology, validation, visualization, writing - original draft, writing -review and editing Lukas Brunner: conceptulization, methodology, validation, visualization, software, writing -review and editing Leonard Borchert: conceptulization, methodology, validation, writing -review and editing Matthew Menary: validation, data curation, software~~

Competing interests. We declare that the authors have no competing interests or conflict of interests.

~~We declare that the authors have no competing interests or conflict of interests.~~

Acknowledgements. This work was carried out as part of the EUCP project which is funded by the European Commission through the Horizon 2020 Programme for Research and Innovation: Grant Agreement 776613.

We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP5 and CMIP6. We thank the climate modeling groups (particularly those listed in tables **??** and ~~SM1~~S1) for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP, CMIP6 and ESGF. Terms of use and further instructions can be found at https://pcmdi.llnl.gov/.

Matthew B. Menary was supported by the European Union Horizon 2020 project 4C, Climate-Carbon Interactions in the Coming Century (grant 821003) and the ANR-Tremplin ERC project HARMONY (ANR-20-ERC9-0001). Leonard Borchert was also funded by the ANR-Tremplin ERC project HARMONY (ANR-20-ERC9-0001) .

References

Athanasiadis, P. J., Ogawa, F., Omrani, N.-E., Keenlyside, N., Schiemann, R., Baker, A. J., Vidale, P. L., Bellucci, A., Ruggieri, P., Haarsma, R., et al.: Mitigating climate biases in the midlatitude North Atlantic by increasing model resolution: SST gradients and their relation to blocking and the jet, Journal of Climate, 35, 3385–3406, 2022.

Bellomo, K., Angeloni, M., Corti, S., and von Hardenberg, J.: Future climate change shaped by inter-model differences in Atlantic meridional overturning circulation response, Nature Communications, 12, 3659, https://doi.org/10.1038/s41467-021-24015-w, 2021.

Bock, L., Lauer, A., Schlund, M., Barreiro, M., Bellouin, N., Jones, C., Meehl, G. A., Predoi, V., Roberts, M. J., and Eyring, V.: Quantifying Progress Across Different CMIP Phases With the ESMValTool, Journal of Geophysical Research: Atmospheres, 125, e2019JD032 321, https://doi.org/10.1029/2019JD032321, 2020.

Booth, B. B. B., Dunstone, N. J., Halloran, P. R., Andrews, T., and Bellouin, N.: Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability, Nature, 484, 228–232, https://doi.org/10.1038/nature10946, 2012.

Borchert, L. F., Pohlmann, H., Baehr, J., Neddermann, N.-C., Suarez-Gutierrez, L., and Müller, W. A.: Decadal Predictions of the Probability of Occurrence for Warm Summer Temperature Extremes, Geophysical Research Letters, 46, 14 042–14 051, https://doi.org/https://doi.org/10.1029/2019GL085385, 2019.

Borchert, L. F., Koul, V., Menary, M. B., Befort, D. J., Swingedouw, D., Sgubin, G., and Mignot, J.: Skillful decadal prediction of unforced southern European summer temperature variations, Environmental Research Letters, 16, 104 017, https://doi.org/10.1088/1748-9326/ac20f5, 2021a.

Borchert, L. F., Menary, M. B., Swingedouw, D., Sgubin, G., Hermanson, L., and Mignot, J.: Improved Decadal Predictions of North Atlantic Subpolar Gyre SST in CMIP6, Geophysical Research Letters, 48, e2020GL091 307, https://doi.org/https://doi.org/10.1029/2020GL091307, 2021b.

Börgel, F., Meier, H. E. M., Gröger, M., Rhein, M., Dutheil, C., and Kaiser, J. M.: Atlantic multidecadal variability and the implications for North European precipitation, Environmental Research Letters, 17, 044 040, https://doi.org/10.1088/1748-9326/ac5ca1, 2022.

Browning, K. A.: The sting at the end of the tail: Damaging winds associated with extratropical cyclones, Quarterly Journal of the Royal Meteorological Society, 130, 375–399, https://doi.org/https://doi.org/10.1256/qj.02.143, 2004.

Brunner, L., Lorenz, R., Zumwald, M., and Knutti, R.: Quantifying uncertainty in European climate projections using combined performance-independence weighting, Environmental Research Letters, 14, 124 010, https://doi.org/10.1088/1748-9326/ab492f, 2019.

Brunner, L., McSweeney, C., Ballinger, A. P., Befort, D. J., Benassi, M., Booth, B., Coppola, E., Vries, H. D., Harris, G., Hegerl, G. C., Knutti, R., Lenderink, G., Lowe, J., Nogherotto, R., O'Reilly, C., Qasmi, S., Ribes, A., Stocchi, P., and Undorf, S.: Comparing Methods to Constrain Future European Climate Projections Using a Consistent Framework, Journal of Climate, 33, 8671–8692, https://doi.org/10.1175/JCLI-D-19-0953.1, 2020a.

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

Carvalho-Oliveira, J., Borchert, L. F., Duchez, A., Dobrynin, M., and Baehr, J.: Subtle influence of the Atlantic Meridional Overturning Circulation (AMOC) on seasonal sea surface temperature (SST) hindcast skill in the North Atlantic, Weather and Climate Dynamics, 2, 739–757, https://doi.org/10.5194/wcd-2-739-2021, 2021.

Carvalho-Oliveira, J., Borchert, L. F., Zorita, E., and Baehr, J.: Self-Organizing Maps Identify Windows of Opportunity for Seasonal European Summer Predictions, Frontiers in Climate, 4, https://doi.org/10.3389/fclim.2022.844634, 2022.

Chen, Z., Zhou, T., Chen, X., Zhang, W., Zhang, L., Wu, M., and Zou, L.: Observationally constrained projection of Afro-Asian monsoon precipitation, Nature Communications, 13, 2552, https://doi.org/10.1038/s41467-022-30106-z, 2022.

D'Andrea, F.: Northern Hemisphere atmospheric blocking as simulated by 15 atmospheric general circulation models in the period 1979–1988, Climate Dyn., 14, 385–407, https://doi.org/10.1007/s003820050230, 1998.

Davini, P. and d'Andrea, F.: From CMIP3 to CMIP6: Northern hemisphere atmospheric blocking simulation in present and future climate, Journal of Climate, 33, 10 021–10 038, https://doi.org/10.1175/JCLI-D-19-0862.1, 2020.

Davini, P., Cagnazzo, C., Gualdi, S., and Navarra, A.: Bidimensional diagnostics, variability and trends of Northern Hemisphere blocking, J. Climate, 25, 6496–6509, https://doi.org/10.1175/JCLI-D-12-00032.1, 2012.

Dong, B., Sutton, R. T., Woollings, T., and Hodges, K.: Variability of the North Atlantic summer storm track: mechanisms and impacts on European climate, Environmental Research Letters, 8, 034 037, https://doi.org/10.1088/1748-9326/8/3/034037, 2013.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., Hurtt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G. A., Pendergrass, A. G., Pincus, R., Ruane, A. C., Russell, J. L., Sanderson, B. M., Santer, B. D., Sherwood, S. C., Simpson, I. R., Stouffer, R. J., and Williamson, M. S.: Taking climate model evaluation to the next level, Nature Climate Change, 9, 102–110, https://doi.org/10.1038/s41558-018-0355-y, 2019.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., and Rummukainen, M.: Evaluation of climate models, pp. 741–882, Cambridge University Press, Cambridge, UK, https://doi.org/10.1017/CBO9781107415324.020, 2013.

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., Lunt, D., Mauritsen, T., Palmer, M., Watanabe, M., Wild, M., and Zhang, H.: The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity, 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, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., pp. 923–1054, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896.009., 2021.

Gervais, M., Shaman, J., and Kushnir, Y.: Impacts of the North Atlantic Warming Hole in Future Climate Projections: Mean Atmospheric Circulation and the North Atlantic Jet, Journal of Climate, 32, 2673–2689, https://doi.org/10.1175/JCLI-D-18-0647.1, 2019.

Gutiérrez, J., Jones, R., Narisma, G., Alves, L., Amjad, M., Gorodetskaya, I., Grose, M., Klutse, N., S.Krakovska, Li, J., Martínez-Castro, D., Mearns, L., Mernild, S., Ngo-Duc, T., van den Hurk, B., and Yoon, J.-H.: Atlas, 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, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., chap. Atlas, pp. 1927–2058, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896.021.1928, 2021.

Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W., and Zelinka, M.: Climate simulations: recognize the 'hot model' problem, Nature, 605, 26–29, https://doi.org/10.1038/d41586-022-01192-2, 2022.

Hempel, S., Frieler, K., Warszawski, L., Schewe, J., and Piontek, F.: A trend-preserving bias correction – the ISI-MIP approach, Earth System Dynamics, 4, 219–236, https://doi.org/10.5194/esd-4-219-2013, 2013.

Hodges, K.: Feature tracking on the unit sphere, Monthly Weather Review, 123, 3458–3465, 1995.

Hodges, K. I.: A General Method for Tracking Analysis and Its Application to Meteorological Data, Monthly Weather Review, 122, 2573 – 2586, https://doi.org/10.1175/1520-0493(1994)122<2573:AGMFTA>2.0.CO;2, 1994.

IPCC: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK and New York, NY, USA, https://www.ipcc.ch/report/ar4/wg1/, 2007.

IPCC: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK and New York, NY, USA, https://www.ipcc.ch/report/ar5/wg1/, 2013.

Ito, R., Shiogama, H., Nakaegawa, T., and Takayabu, I.: Uncertainties in climate change projections covered by the ISIMIP and CORDEX model subsets from CMIP5, Geoscientific Model Development, 13, 859–872, https://doi.org/10.5194/gmd-13-859-2020, 2020.

Jackson, L. C., Biastoch, A., Buckley, M. W., Desbruyères, D. G., Frajka-Williams, E., Moat, B., and Robson, J.: The evolution of the North Atlantic Meridional Overturning Circulation since 1980, Nature Reviews Earth & Environment, 3, 241–254, https://doi.org/10.1038/s43017-022-00263-2, 2022.

Jin, C., Wang, B., and Liu, J.: Future Changes and Controlling Factors of the Eight Regional Monsoons Projected by CMIP6 Models, Journal of Climate, 33, 9307–9326, https://doi.org/10.1175/JCLI-D-20-0236.1, 2020.

Kaspi, Y. and Schneider, T.: The Role of Stationary Eddies in Shaping Midlatitude Storm Tracks, Journal of the Atmospheric Sciences, 70, 2596–2613, https://doi.org/10.1175/JAS-D-12-082.1, 2013.

Keeley, S. P. E., Sutton, R. T., and Shaffrey, L. C.: The impact of North Atlantic sea surface temperature errors on the simulation of North Atlantic European region climate, Quarterly Journal of the Royal Meteorological Society, 138, 1774–1783, https://doi.org/https://doi.org/10.1002/qj.1912, 2012.

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

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, Geophysical Research Letters, 44, 1909–1918, https://doi.org/https://doi.org/10.1002/2016GL072012, 2017.

Lange, S. and Büchner, M.: ISIMIP3b bias-adjusted atmospheric climate input data (v1.1), Tech. rep., https://doi.org/https://doi.org/10.48364/ISIMIP.842396.1, 2021.

Lee, R. W., Woollings, T. J., Hoskins, B. J., Williams, K. D., O'Reilly, C. H., and Masato, G.: Impact of Gulf Stream SST biases on the global atmospheric circulation, Climate Dynamics, 51, 3369–3387, https://doi.org/10.1007/s00382-018-4083-9, 2018.

Liang, Y., Gillett, N. P., and Monahan, A. H.: Climate Model Projections of 21st Century Global Warming Constrained Using the Observed Warming Trend, Geophysical Research Letters, 47, e2019GL086 757, https://doi.org/https://doi.org/10.1029/2019GL086757, 2020.

Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., and Knutti, R.: Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections Over North America, Journal of Geophysical Research: Atmospheres, 123, 4509–4526, https://doi.org/https://doi.org/10.1029/2017JD027992, 2018.

Lutz, A. F., ter Maat, H. W., Biemans, H., Shrestha, A. B., Wester, P., and Immerzeel, W. W.: Selecting representative climate models for climate change impact studies: an advanced envelope-based selection approach, International Journal of Climatology, 36, 3988–4005, https://doi.org/https://doi.org/10.1002/joc.4608, 2016.

Mahony, C. R., Wang, T., Hamann, A., and Cannon, A. J.: A global climate model ensemble for downscaled monthly climate normals over North America, International Journal of Climatology, 42, 5871–5891, https://doi.org/https://doi.org/10.1002/joc.7566, 2022.

Masato, G., Hoskins, B. J., and Woollings, T.: Winter and summer Northern Hemisphere blocking in CMIP5 models, J. Climate, 26, 7044–7059, https://doi.org/10.1175/JCLI-D-12-00466.1, 2013.

McDermid, S. P., Ruane, A. C., Rosenzweig, C., Hudson, N. I., Morales, M. D., Agalawatte, P., Ahmad, S., Ahuja, L. R., Amien, I., Anapalli, S. S., Anothai, J., Asseng, S., Biggs, J., Bert, F., Bertuzzi, P., Bhatia, V. S., Bindi, M., Broad, I., Cammarano, D., Carretero, R., Chattha, A. A., Chung, U., Debats, S., Deligios, P., De Sanctis, G., Dhliwayo, T., Dumont, B., Estes, L., Ewert, F., Ferrise, R., Gaiser, T., Garcia, G., Gbegbelegbe, S., Geethalakshmi, V., Gerardeaux, E., Goldberg, R., Grant, B., Guevara, E., Hickman, J., Hoffmann, H., Huang, H., Hussain, J., Justino, F. B., Karunaratne, A. S., Koehler, A.-K., Kouakou, P. K., Kumar, S. N., Lakshmanan, A., Lieffering, M., Lin, X., Luo, Q., Magrin, G., Mancini, M., Marin, F. R., Marta, A. D., Masutomi, Y., Mavromatis, T., McLean, G., Meira, S., Mohanty, M., Moriondo, M., Nasim, W., Negm, L., Orlando, F., Orlandini, S., Ozturk, I., Soares Pinto, H. M., Podesta, G., Qi, Z., Ramarohetra, J., ur Rahman, M. H., Raynal, H., Rodriguez, G., Rötter, R., Sharda, V., Shuo, L., Smith, W., Snow, V., Soltani, A., Srinivas, K., Sultan, B., Swain, D. K., Tao, F., Tesfaye, K., Travasso, M. I., Trombi, G., Topaj, A., Vanuytrecht, E., Viscarra, F. E., Aftab Wajid, S., Wang, E., Wang, H., Wang, J., Wijekoon, E., Byun-Woo, L., Xiaoguang, Y., Young, B. H., Yun, J. I., Zhao, Z., and Zubair, L.: The AgMIP Coordinated Climate-Crop Modeling Project (C3MP): Methods and Protocols, in: Handbook of Climate Change and Agroecosystems, vol. Volume 3 of *ICP Series on Climate Change Impacts, Adaptation, and Mitigation*, pp. 191–220, IMPERIAL COLLEGE PRESS, https://doi.org/doi:10.1142/9781783265640_0008, 2014.

McSweeney, C., Murphy, J., Sexton, D., Rostron, J., Yamazaki, K., and Harris, G.: Selection of CMIP5 members to augment a perturbed–parameter ensemble of global realisations of future climate for the UKCP18 scenarios., Tech. rep., Hadley Centre Technical Note 102. HCTN_102_2018P | Met Office UA, 2018.

McSweeney, C. F. and Jones, R. G.: How representative is the spread of climate projections from the 5 CMIP5 GCMs used in ISI-MIP?, Climate Services, 1, 24–29, https://doi.org/https://doi.org/10.1016/j.cliser.2016.02.001, 2016.

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

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

Menary, M. B., Robson, J., Allan, R. P., Booth, B. B. B., Cassou, C., Gastineau, G., Gregory, J., Hodson, D., Jones, C., Mignot, J., Ringer, M., Sutton, R., Wilcox, L., and Zhang, R.: Aerosol-Forced AMOC Changes in CMIP6 Historical Simulations, Geophysical Research Letters, 47, https://doi.org/10.1029/2020GL088166, 2020.

Merrifield, A. L., Brunner, L., Lorenz, R., Medhaug, I., and Knutti, R.: An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles, Earth System Dynamics, 11, 807–834, https://doi.org/10.5194/esd-11-807-2020, 2020.

Michelangeli, P.-A., Vautard, R., and Legras, B.: Weather Regimes: Recurrence and Quasi Stationarity, Journal of Atmospheric Sciences, 52, 1237–1256, https://doi.org/10.1175/1520-0469(1995)052<1237:WRRAQS>2.0.CO;2, 1995.

MS, B. and JA, T.: Weighting a regional climate model ensemble: Does it make a difference? Can it make a difference? , Climate Research, 77, 23–43, https://www.int-res.com/abstracts/cr/v77/n1/p23-43/, 2019.

Ossó, A., Sutton, R., Shaffrey, L., and Dong, B.: Development, Amplification, and Decay of Atlantic/European Summer Weather Patterns Linked to Spring North Atlantic Sea Surface Temperatures, Journal of Climate, 33, 5939–5951, https://doi.org/10.1175/JCLI-D-19-0613.1, 2020.

Oudar, T., Cattiaux, J., and Douville, H.: Drivers of the Northern Extratropical Eddy-Driven Jet Change in CMIP5 and CMIP6 Models, Geophysical Research Letters, 47, e2019GL086 695, https://doi.org/https://doi.org/10.1029/2019GL086695, 2020.

Overland, J. E., Wang, M., Bond, N. A., Walsh, J. E., Kattsov, V. M., and Chapman, W. L.: Considerations in the Selection of Global Climate Models for Regional Climate Projections: The Arctic as a Case Study, Journal of Climate, 24, 1583–1597, https://doi.org/10.1175/2010JCLI3462.1, 2011.

Palmer, T. E., Booth, B. B. B., and McSweeney, C. F.: How does the CMIP6 ensemble change the picture for European climate projections?, Environmental Research Letters, 16, 094 042, https://doi.org/10.1088/1748-9326/ac1ed9, 2021.

Pelly, J. and Hoskins, B.: A new perspective on blocking, J. Atmos. Sci., 60, 743–755, https://doi.org/10.1175/1520-0469(2003)060<0743:ANPOB>2.0.CO;2, 2003.

Prein, A. F., Bukovsky, M. S., Mearns, L. O., Bruyère, C. L., and Done, J. M.: Simulating North American Weather Types With Regional Climate Models , https://www.frontiersin.org/article/10.3389/fenvs.2019.00036, 2019.

Priestley, M. D., Ackerley, D., Catto, J. L., Hodges, K. I., McDonald, R. E., and Lee, R. W.: An Overview of the Extratropical Storm Tracks in CMIP6 Historical Simulations, Journal of Climate, 33, 6315–6343, https://doi.org/10.1175/JCLI-D-19-0928.1, 2020.

Priestley, M. D. K., Ackerley, D., Catto, J. L., and Hodges, K. I.: Drivers of biases in the CMIP6 extratropical storm tracks. Part 1: Northern Hemisphere, Journal of Climate, pp. 1–37, https://doi.org/10.1175/jcli-d-20-0976.1, 2022.

Priestley, M. D. K., Ackerley, D., Catto, J. L., and Hodges, K. I.: Drivers of Biases in the CMIP6 Extratropical Storm Tracks. Part I: Northern Hemisphere, Journal of Climate, 36, 1451–1467, https://doi.org/10.1175/JCLI-D-20-0976.1, 2023.

Rex, D.: Blocking action in the middle troposphere and its effect upon regional climate: I. An aerological study of blocking action, Tellus, 2, 196–211, 1950.

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

Ribes, A., Boé, J., Qasmi, S., Dubuisson, B., Douville, H., and Terray, L.: An updated assessment of past and future warming over France based on a regional observational constraint, Earth System Dynamics, 13, 1397–1415, https://doi.org/10.5194/esd-13-1397-2022, 2022.

Rosenzweig, C., Arnell, N. W., Ebi, K. L., Lotze-Campen, H., Raes, F., Rapley, C., Smith, M. S., Cramer, W., Frieler, K., Reyer, C. P., Schewe, J., Van Vuuren, D., and Warszawski, L.: Assessing inter-sectoral climate change risks: The role of ISIMIP, Environmental Research Letters, 12, 010 301, https://doi.org/10.1088/1748-9326/12/1/010301, 2017.

Ruane, A. C. and McDermid, S. P.: Selection of a representative subset of global climate models that captures the profile of regional changes for integrated climate impacts assessment, Earth Perspectives, 4, 1, https://doi.org/10.1186/s40322-017-0036-4, 2017.

Ruane, A. C., McDermid, S., Rosenzweig, C., Baigorria, G. A., Jones, J. W., Romero, C. C., and DeWayne Cecil, L.: Carbon–Temperature–Water change analysis for peanut production under climate change: a prototype for the AgMIP Coordinated Climate-Crop Modeling Project (C3MP), Global Change Biology, 20, 394–407, https://doi.org/https://doi.org/10.1111/gcb.12412, 2014.

Scaife, A. A., Copsey, D., Gordon, C., Harris, C., Hinton, T., Keeley, S., O'Neill, A., Roberts, M., and Williams, K.: Improved Atlantic winter blocking in a climate model, Geophysical Research Letters, 38, https://doi.org/https://doi.org/10.1029/2011GL049573, 2011.

Schiemann, R., Athanasiadis, P., Barriopedro, D., Doblas-Reyes, F., Lohmann, K., Roberts, M. J., Sein, D. V., Roberts, C. D., Terray, L., and Vidale, P. L.: Northern Hemisphere blocking simulation in current climate models: evaluating progress from the Climate Model Intercom-

parison Project Phase 5 to 6 and sensitivity to resolution, Weather and Climate Dynamics, 1, 277–292, https://doi.org/10.5194/wcd-1-277-2020, 2020.

Selten, F. M., Bintanja, R., Vautard, R., and van den Hurk, B. J. J. M.: Future continental summer warming constrained by the present-day seasonal cycle of surface hydrology, Scientific Reports, 10, 4721, https://doi.org/10.1038/s41598-020-61721-9, 2020.

Shepherd, T. G.: Atmospheric circulation as a source of uncertainty in climate change projections, Nature Geoscience, 7, 703–708, https://doi.org/10.1038/ngeo2253, 2014.

Shepherd, T. G.: Storyline approach to the construction of regional climate change information, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 475, 20190 013, https://doi.org/10.1098/rspa.2019.0013, 2019.

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., Watanabe, M., Andrews, T., Braconnot, P., Bretherton, C. S., Foster, G. L., Hausfather, Z., von der Heydt, A. S., Knutti, R., Mauritsen, T., Norris, J. R., Proistosescu, C., Rugenstein, M., Schmidt, G. A., Tokarska, K. B., and Zelinka, M. D.: An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence, Reviews of Geophysics, 58, https://doi.org/10.1029/2019RG000678, 2020.

Shiogama, H., Ishizaki, N. N., Hanasaki, N., Takahashi, K., Emori, S., Ito, R., Nakaegawa, T., Takayabu, I., Hijioka, Y., Takayabu, Y. N., and Shibuya, R.: Selecting CMIP6-Based Future Climate Scenarios for Impact and Adaptation Studies, SOLA, 17, 57–62, https://doi.org/10.2151/sola.2021-009, 2021.

Simpson, I. R., Deser, C., McKinnon, K. A., and Barnes, E. A.: Modeled and Observed Multidecadal Variability in the North Atlantic Jet Stream and Its Connection to Sea Surface Temperatures, Journal of Climate, 31, 8313–8338, https://doi.org/10.1175/JCLI-D-18-0168.1, 2018.

Sutton, R. T. and Dong, B.: Atlantic Ocean influence on a shift in European climate in the 1990s, Nature Geoscience, 5, 788–792, https://doi.org/10.1038/ngeo1595, 2012.

Tibaldi, S. and Molteni, F.: On the operational predictability of blocking, Tellus, 42A, 343–365, https://doi.org/10.3402/tellusa.v42i3.11882, 1990.

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

Tsujino, H., Urakawa, L. S., Griffies, S. M., Danabasoglu, G., Adcroft, A. J., Amaral, A. E., Arsouze, T., Bentsen, M., Bernardello, R., Böning, C. W., Bozec, A., Chassignet, E. P., Danilov, S., Dussin, R., Exarchou, E., Fogli, P. G., Fox-Kemper, B., Guo, C., Ilicak, M., Iovino, D., Kim, W. M., Koldunov, N., Lapin, V., Li, Y., Lin, P., Lindsay, K., Liu, H., Long, M. C., Komuro, Y., Marsland, S. J., Masina, S., Nummelin, A., Rieck, J. K., Ruprich-Robert, Y., Scheinert, M., Sicardi, V., Sidorenko, D., Suzuki, T., Tatebe, H., Wang, Q., Yeager, S. G., and Yu, Z.: Evaluation of global ocean–sea-ice model simulations based on the experimental protocols of the Ocean Model Intercomparison Project phase 2 (OMIP-2), Geoscientific Model Development, 13, 3643–3708, https://doi.org/10.5194/gmd-13-3643-2020, 2020.

van den Hurk, B., Siegmund, P., Klien Tank (Eds), A., Attema, J., Bakker, A., Beersma, J., Bessembinder, J., Boers, R., Brandsma, T., van de Brink, H., Drijfhout, S., Eskes, H., Haarsma, R., Hazeleger, W., Jilderda, R., Katsman, C., Lenderink, G., Loriaux, J., van de Meijgaard, E., van Noije, T., van Oldenborgh, G. J., Selten, F., Siebesma, P., Sterl, A., de Vries, H., Van Weele, M., de Winter, R., and van Zadelhoff, G.-J.: KNMI'14: Climate Change scenarios for the 21st Century – A Netherlands perspective, Tech. rep., Royal Netherlands Meteorological Istitute Ministry of Infrastructure and Water Management, ttps://www.knmiprojects.nl/projects/climate-scenarios, 2014.

Whetton, P., Macadam, I., Bathols, J., and O'Grady, J.: Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models, Geophysical Research Letters, 34, https://doi.org/https://doi.org/10.1029/2007GL030025, 2007.

White, J. W., Hoogenboom, G., Kimball, B. A., and Wall, G. W.: Methodologies for simulating impacts of climate change on crop production, Field Crops Research, 124, 357–368, https://doi.org/https://doi.org/10.1016/j.fcr.2011.07.001, 2011.

Yeager, S. G. and Robson, J. I.: Recent Progress in Understanding and Predicting Atlantic Decadal Climate Variability, Current Climate Change Reports, 3, 112–127, https://doi.org/10.1007/s40641-017-0064-z, 2017.

Zappa, G. and Shepherd, T. G.: Storylines of atmospheric circulation change for European regional climate impact assessment, Journal of Climate, https://doi.org/10.1175/JCLI-D-16-0807.1, 2017.

Zappa, G., Shaffrey, L. C., and Hodges, K. I.: The ability of CMIP5 models to simulate North Atlantic extratropical cyclones, J. Climate, 26, 5379–5396, https://doi.org/10.1175/JCLI-D-12-00501.1, 2013.

Zhang, M.-Z., Xu, Z., Han, Y., and Guo, W.: Evaluation of CMIP6 models toward dynamical downscaling over 14 CORDEX domains, Climate Dynamics, https://doi.org/10.1007/s00382-022-06355-5, 2022.

Zhang, R.: Coherent surface-subsurface fingerprint of the Atlantic meridional overturning circulation, Geophysical Research Letters, 35, https://doi.org/10.1029/2008GL035463, 2008.

Zhang, R., Sutton, R., Danabasoglu, G., Kwon, Y., Marsh, R., Yeager, S. G., Amrhein, D. E., and Little, C. M.: A Review of the Role of the Atlantic Meridional Overturning Circulation in Atlantic Multidecadal Variability and Associated Climate Impacts, Reviews of Geophysics, 57, 316–375, https://doi.org/10.1029/2019RG000644, 2019.