Copernicus Climate Change Service

Documentation for the role of internal variability over Europe

D34b_Lot2.1.3.2

Issued by: SMHI/ Anna Eronn

Date: Originally 31/12/2020; postponed to 31/3/2021

Ref: C3S_ D34b_Lot2.1.3.2_202103_ Documentation for the role of internal variability over Europe_v5

IMPLEMENTED BY ECMWF

Contributors

DMI

Ole B. Christensen

KNMI

Erik van Meijgaard, Emma Aalbers

SMHI

Grigory Nikulin

Erik Kjellström

ETHZ

Marie-Estelle Demory

CNRS

Robert Vautard

MF

Samuel Somot

ICTP

Erika Coppola

MOHC

Erasmo Buonomo,

Richard Jones

HZG

Claas Teichmann

Table of Contents

1. Executive Summary

This report is Deliverable C3S_D34b_Lot2.1.3.2 of the COPERNICUS C3S_D34b_Lot2 (PRINCIPLES) project; it describes regional and local effects of internal climate variability based on several model ensembles: two sets of downscaled data for two different three-member GCM ensembles; and two downscaled single-model multi-member ensembles available from other projects. The plan for parts of the analysis has already been laid out in C3S_D34b_Lot2.1.3.1, and this report aims at describing the results.

Five PRINCIPLES partner institutions have used their respective RCMs to downscale three ensemble members of one or two GCMs generating a total of 21 simulations studied as two separate simulation matrices.

This report contains two separate analyses of internal variability: An analysis of grid-scale and regional-scale responses for various climate indices, and a specific ANOVA-based significance analysis of the roles of ensemble member and RCM model on temperature, precipitation, and wind and their change.

A full domain grid-scale analysis of climate change responses shows that the internal variability (inter-member spread) at 30-year time scales in mean sea level pressure, being a measure of large-scale circulation, is much larger than the inter-RCM-model differences (uncertainties coming from the model formulations). Inter-member spread related to circulation variability is likewise seen in the change of high- and low-end percentiles of temperature and in the relative change in mean precipitation, but for these parameters the inter-model spread is of comparable size. For extreme precipitation, both the inter-member spread and the inter-model spread are larger than for the other parameters, and, in particular for JJA, large variations are apparent at small spatial scales. Stratified on climate index the grid-scale climate change signal for all three-member ensembles varies from nearly everywhere significant for extreme temperature to virtually nowhere significant for extreme precipitation in summer.

Results from a catchment-scale analysis focusing on the temporal evolution of the response show there are large inter-member differences owing to long-term fluctuations throughout the full period of simulation. These differences can be related to the role of a driving GCM member or to the behaviour of specific GCM-RCM combinations. Much smoother response time series are found for two larger single-model multi-member ensembles, available from other projects, but small fluctuations remain visible. In fact, climate change signals derived from randomly composed subsamples can evolve outside the full ensemble spread envelope during prolonged time intervals. This demonstrates the importance of using multi-member ensembles for an accurate estimation of the climate change response, even at the scale of a large river catchment.

Regarding the combination of temperature and precipitation, the reference state of the mean climate in both winter and summer shows considerable inter-model differences which are larger than the inter-member differences. For the future combined change in mean temperature and precipitation the inter-member spread has a more prominent role. In both downscaled three-member GCM ensembles the inter-model spread in summer temperature response for the end-of-century period is larger than the inter-member spread, implying that uncertainty due to model differences has become larger than uncertainty due to internal variability.

The ANOVA analysis shows that the GCM ensemble member has a significant effect on the mean state of the climate for a 30-year time slice for large parts of the integration area for wind speed and precipitation and especially for temperature. We conclude that an ensemble size of three GCM simulations is too small to get a good signal-to-noise ratio for the mean climate. The mean climate has a robust dependence on the RCM almost everywhere for all fields studied.

Regarding climate change, temperature shows the largest effect of long-term fluctuations as reflected in quite large areas of significant differences between ensemble members of the same GCM. The RCM model is mostly insignificant but has, however, a large influence over parts of the North Atlantic for winter temperature change as well as over most of the continent for summer temperature change. In winter the areas with the largest climate change variability among GCM-ensemble members are located in south-eastern Europe and in northern Scandinavia, but this variability is smaller than interannual variability; in summer, France and Spain as well as parts of Scandinavia and south-eastern Europe show the largest ensemble-member dependence of climate change.

Internal, natural, variability is a prominent feature of the climate that can give rise to strong differences between time periods even without changes in forcing conditions. Large ensembles with many climate model simulations can sample such differences and thereby be used to assess what are robust features of climate change and what is more difficult to detect out of the background noise. We find that such differences are important for the European climate, both for the large-scale atmospheric circulation and for local conditions, including extreme events. The results emphasize that large multi-model ensembles should be used for assessing climate change on local and regional scales.

The over-all results from both analyses indicate that, although there can be distinct inter-model differences in the representation of the climate and climate change, the size of the inter-member spread appears to be rather insensitive to the model specifics, and is much more determined by the climate parameter that is examined. At the same time, three-member ensembles are, in general, too small to generate credible maps of inter-member spread. These findings indicate that downscaling a selection of single-model multi-member GCMs with a single RCM or a few RCMs may be enough to yield an accurate estimate of the inter-member spread that is representative for RCMs in general. It would therefore be relevant in the design of a future multi-model ensemble to include some single-GCM sub-ensembles as part of the multi-model ensemble, in particular for the assessment of changes in extremes.

2. Introduction

The PRINCIPLES simulations cover the Euro-CORDEX EUR-11 area at 12.5km horizontal resolution (http://www.cordex.org). It is a novel feature that several different regional models have downscaled the same sets of single-GCM ensembles and it is therefore the first time that we have a chance to separate RCM effects from downscaled internal variability. In this report we document how the issue of internal model variability is approached in the PRINCIPLES project.

The variability of climate model output mainly originates from three quite different sources: emission scenario uncertainty, model differences, and climate variability as simulated by models. In this study we focus on the third of these sources of variability; one approach in this report is that we aim to

separate the GCM internal variability from RCM model variability through an ANOVA analysis, in order to extract the maximum possible information from the available simulation output.

The available data consist of downscaling simulations with different regional models, two different sets for the two GCM mini-ensembles analysed; the three-member ICHEC-EC-EARTH ensemble has been downscaled with 4 different RCMs, whereas the three-member MPI-M-MPI-ESM-LR has been downscaled with 3 RCMs.

The PRINCIPLES project aimed at defining an experimental protocol that allowed assessing how internal climate variability represented by models affects estimates of climate change and uncertainties of such estimates. Two sources of internal variability are investigated; any individual downscaling simulation will contain both components indistinguishably: variability attributable to the driving GCM (e.g. due to decadal or longer variability and to large-scale weather systems) or to the RCM (e.g. different systematic fine-scale behaviour depending on the atmospheric state). For this purpose we identify an appropriate set of ensemble members from one or several GCMs to be downscaled by the RCMs in PRINCIPLES, and we assess internal climate variability based on the new simulations resulting from the project as well as from already existing additional information.

Large single-model multi-member ensembles have been studied before. Deser et al. (2012) examined a 40-member ensemble of the CCSM3 global model in terms of e.g. the decade of emergence of statistical significance of various climate signals and quantified the influence of internal variability on the significance of the climate change signal. Maher et al. (2021) used six single-model initial condition large ensembles (SMILEs) to separate the (reducible) uncertainty in the forced response due to model-to-model differences from the (irreducible) uncertainty due to internal variability and concluded that, on a global scale, the former dominates in the projected changes of mean temperature and precipitation, while the latter mostly dominates in projected changes of temporal variability of temperature and precipitation. Milinski et al. (2020) introduced an objective method to estimate the ensemble size that is required to separate the forced response from the internal variability dependent on the type of application and the amount of error the user is willing to accept.

Also downscaling of single-model ensembles has been performed. Addor and Fischer (2014) downscaled a single-model transient-run GCM ensemble of 21 members of the NCAR CESM model with the COSMO-CLM regional climate model at 50 km resolution in order to quantify the role of internal variability on transient climate change over the Alps. Aalbers et al. (2017) found that with the 16-member KNMI-ensemble extreme precipitation, which has very large natural variability, can be analysed in a much more robust way at the local scale than would have been possible with just one simulation. Potential applications of single-model multi-member ensembles include: representing multi-model spread through resampling multi-member spread (Lenderink et al., 2014); compilation of very long multiple-parameter time series to facilitate the study of compound events (van den Hurk et al., 2015); and the use of very long time series to estimate the occurrence of wind extremes (Vautard et al., 2019) and precipitation extremes (Philip et al., 2018). In recent studies, von Trentini et al. (2018; 2019) compared natural variability in the 50-member ClimEx-ensemble with a 22 member multi-model ensemble compiled of a subset of EURO-CORDEX simulations, and assessed the inter-annual variability in the three above-mentioned initial-condition multi-member RCM ensembles.

In this report we focus on the downscaled internal variability and assess how the spread in weather modes originating from global models will manifest itself in the high-resolution signals simulated by RCMs. This will involve more than a single regional model in order to study the relative effects of internal variability and model choice directly. As the current ensemble uses several RCMs and two separate sets of downscaled GCM ensemble members, we have an unprecedented opportunity to study common features, which do not depend on the RCM model or the GCM model.

This will be done through a targeted analysis of internal variability for a few of the available model combinations in the larger GCM-RCM-scenario matrix under consideration in PRINCIPLES by performing regional downscaling of several ensemble models of selected GCMs. Using ensembles of transient simulations will allow quantification of climate for selected periods with a relatively high degree of accuracy by generating a large sample of simulated years in the periods under investigation.

In PRINCIPLES, the complete set of simulations builds on an already existing limited set in order to achieve matrices with as high a degree of filling as possible. Therefore, a set of six CMIP5 GCM models have been chosen on the grounds that they have been downscaled several times already at the beginning of this project. These six models, which have all been used to simulate the relevant set of emission scenarios, are: HadGEM2-ES, EC-Earth, CNRM-CM5, NorESM1-M, MPI-ESM-LR, and IPSL-CM5A-MR. Out of these six models, unfortunately, only two have saved the data necessary for generating RCM boundary files for at least three ensemble members: EC-Earth and MPI-ESM-LR. Therefore, three ensemble members of each of these two GCMs have been downscaled with several of the project RCMs. The choice of simulations can be seen in Fig. 1.

The number of CMIP5 GCM simulations where sufficient boundary data have been saved is quite high, of the order of 30 (e.g., McSweeney et al., 2014). However, due to the aim of filling as much of the matrix as possible, we have prioritized consistency with prior simulations. Without this constraint a more optimal choice, based on GCM and RCM performance in simulating the historical climate and simulating a wide plausible range of future climate change, could have been made and it will be important to situate these results into the broader context of the full available ensemble. This also means that we have only chosen among GCMs with pre-existing simulations for this particular study of internal variability.

The mini-ensembles are also analysed over a part of the Rhine river catchment and the results are compared to two larger ensembles: a 16-member GCM (EC-Earth) ensemble generated by KNMI, which has been downscaled with one of the RCMs (RACMO-22E) participating in PRINCIPLES (Aalbers et al. 2017), and a 50-member CanESM2-CRCM5 ensemble, produced within the ClimEx-project (Leduc et al., 2019). Both multi-member ensembles employed the same 12-km resolution as the PRINCIPLES simulations. The RCM-domain of the KNMI-ensemble covers Western Europe, while the downscaling domain within ClimEx coincides with a substantial part of the CORDEX-EUR-011 domain.

3. Method

In order to have a reasonable number of RCMs to investigate internal variability, five of the nine institutions participating in PRINCIPLES (see Table 1) have performed one or two sets of downscaling of three-member GCM ensembles. We will concentrate on one emission scenario, the RCP8.5, which has the highest pre-project population of downscaling simulations.

The total number of single-scenario simulations in PRINCIPLES is 68. For the study of internal variability, we have chosen to use 18 simulations. As seen in Fig. 1, we have two GCMs, each contributing 3 ensemble members, each of which downscaled with 3 RCMs. The missing simulation has been emulated for the ANOVA analysis with the method described in deliverable C3S_D34b_Lot2.1.4.1. The RCMs are not the same set for the two GCM matrices. We shall refer to the EC-EARTH-driven matrix as M1, and to the MPI-ESM-driven matrix as M2.

Table 1. PRINCIPLES institutes and regional models. The models in bold typeface are taking part in the specific investigation of internal variability presented here.

Number	Partner institute	Regional Climate Model
1	**SMHI**	**RCA4**
2	**ETHZ**	**CrCLM**
3	**HZG**	**REMO2009 and REMO2015**
4	**KNMI**	**RACMO22E**
5	**DMI**	**HIRHAM5**
6	CNRS-IPSL	WRF381P
7	Météo-France	ALADIN63
8	OGS/ICTP	RegCM4.6.1
9	MOHC	HadGEM3-RA

	SMHI RCA4	ETHZ CrCLM	HZG REMO	KNMI RACMO22E	DMI HIRHAM5
ICHEC-EC-EARTH	3	3	1	3	3
MPI-M-MPI-ESM-LR	3	3	3	1	1

Figure 1. The GCMxRCM sub-matrix where internal GCM model variability is investigated. The light coloured areas contain one simulation, and the dark grey areas contain downscaled simulations of three GCM ensemble members. Numbers indicate the number of existing simulations with the GCM, RCM in question at the time of this report, 03/2021. The only emission scenario considered is RCP8.5.

The analysis consists of two main steps. An analysis of several specific derived fields for each member of the PRINCIPLES ensemble, where EC-EARTH is the driving GCM (3 GCM ensemble members, each downscaled with the same 4 RCMs), including a comparison for a part of the Rhine river catchment

with other multi-member single-GCM simulations. In this case, significance of climate change for the fields is studied for each RCM and then compared among RCMs. Then, an analysis-of-variance (ANOVA) technique is applied to average temperature, precipitation, and wind speed for the two PRINCIPLES 12- or 9-member ensembles in order to assign the spread between simulations to ensemble member and RCM effects in a simultaneous analysis, aiming at an identification of areas and fields where the choice of GCM ensemble member makes a difference.

4. Results and Discussion

4.1 Analysis of multi-member ensembles

4.1.1 Internal variability of derived fields

In this section the role of internal variability is illustrated for the two GCM drivers (M1=EC-EARTH and M2=MPI-ESM-LR) separately (Fig. 1). Each GCM contributes with three realizations (those that provided boundary conditions for dynamical downscaling; labelled r1, r3 and r12 for M1; r1, r2 and r3 for M2), and each realization is downscaled by different RCMs, four for M1 and three for M2. In this way two multi-RCM multi-GCM-member ensembles have been generated. Since the GCM is the primary driver of each system, the results hereafter are stratified along GCM. Results are shown for the winter (DJF) and the summer season (JJA), and for a limited number of climate indices that are considered of general interest: for mean sea level pressure (psl) the mean is shown for both seasons, for near-surface air temperature (tas) the coldest day in DJF and the warmest day in JJA, and for precipitation (pr) the mean for both seasons, the 5-day maximum amount in DJF and the 1-day maximum amount in JJA. The focus is on the difference between the end-of-century period (2071-2100) and a reference period, here 1981-2010. In the transient simulations the historical emission scenario is used until 2005, and the future scenario RCP8.5 from 2006 onward.

The response in mean sea level pressure in the M1-driven ensemble is shown in Fig. 2. Obviously, there is a considerable difference in DJF-response among the three realizations, with a much larger positive change over the northern part of the domain in member r1 compared to r12 and, in particular, r3. A similarly sized difference is seen for JJA when the amplification of high pressure over the Atlantic west of the British Isles is much stronger in r12 and r1 compared to r3. As a result of these large differences the signal-to-noise ratio (last row in the figures) in both seasons is only significant over relatively small portions of the EUR-11 domain. (The response in a given parameter is significantly different from zero at the 5% significance level for $|S/N| >= \tau$, based on a two-sided t-test. For 3 members, implying 2 degrees of freedom, $\tau = 3.04$. This value is used in displaying the normalized signal-to-noise (S/N) ratio; see (S/N)/τ legend in Figures 2-5, and Figures A1-A4).

The response in mean sea level pressure in the M2-driven ensemble is shown in Fig A1. The main finding for the M1-driven ensemble, namely that there are considerable differences among the different realizations irrespective of the season, is seen in the M2-driven ensemble as well, though there are of course differences in the details. Similarly large differences in long-term psl-response were already reported by Deser et al. (2012) from analyzing a 40-member single-model ensemble. They explored potential reasons for the large inter-member spread and concluded that, at least in the extra-tropics, this was to a large extent driven by internal atmospheric variability induced by intrinsic atmospheric dynamics. Almost half a century ago, Madden (1976) identified such internal

natural variability of monthly means as the variability that results from variance and autocorrelation related to daily weather fluctuations and phrased this as a measure of “climatic noise” within an “unchanging climate”. The effect on the spread in the responses is most profound on mean sea level pressure, but the internal atmospheric variability also substantially contributes to the long-term spread in the response of precipitation and temperature.

The expectation that the synoptic-scale circulation is to a large extent determined by the driving GCM, irrespective of the employed RCM, is confirmed by the inter-RCM-spread for each of the members (last column in Fig. 2) being much smaller than the inter-member spread for each of the RCMs (5th row in Fig. 2). For the same reason, there is a large amount of correspondence between the response of any of the four RCMs driven by a given GCM-member (columns 2 to 5) and the response of the GCM-realization itself (column 1).

The response in the cold day (P05) mean temperature for DJF in the M1-driven ensemble is shown in Fig. 3 as is the response in the warm day (P95) mean temperature for JJA. For these parameters the role of the RCMs is more distinctive than for psl. While the cold-day response in winter is most prominent in Northern Europe, the warm-day response is largest in the Mediterranean and South-western Europe. Spatially averaged, the response in mean temperature of cold winter days is 1.0 (GCM) through 1.6 (RACMO2), 1.7 (HIRHAM5), 2.0 (crCLIM), to 2.2 (RCA4) degrees larger than the response in mean temperature of warm summer days. For P05-tas in DJF the RCMs provide a greater response than the driving GCM, which is likely related, at least partly, to the finer resolution treatment of processes involving snow in the RCMs; for P95-tas in JJA the RCM-response is somewhat smaller than or generally equal (RCA4) to the GCM-response. Also there is more smaller-scale structure related to the RCMs (for example, see the different responses over France among the three RCMs, both in DJF and JJA). On the other hand, the large-scale structure is relatively similar to the response in the driving GCM, albeit responses in Eastern Europe projected by crCLIM and, to a lesser extent, HIRHAM5 are profoundly smaller. The inter-RCM spread is comparable or slightly larger than the inter-member spread. The signal-to-noise ratio is (almost) everywhere (far) beyond τ, which for temperature parameters does not come as a surprise. An exception to this is the crCLIM-response in P95-tas for JJA over the Baltic States, Belarus, and parts of Poland, Russia and Ukraine. The potential reason for this is discussed in the next section and in the Summary.

Results for M2 temperature are shown in Fig. A2, showing similar spreads as seen for M1 in Fig. 3.

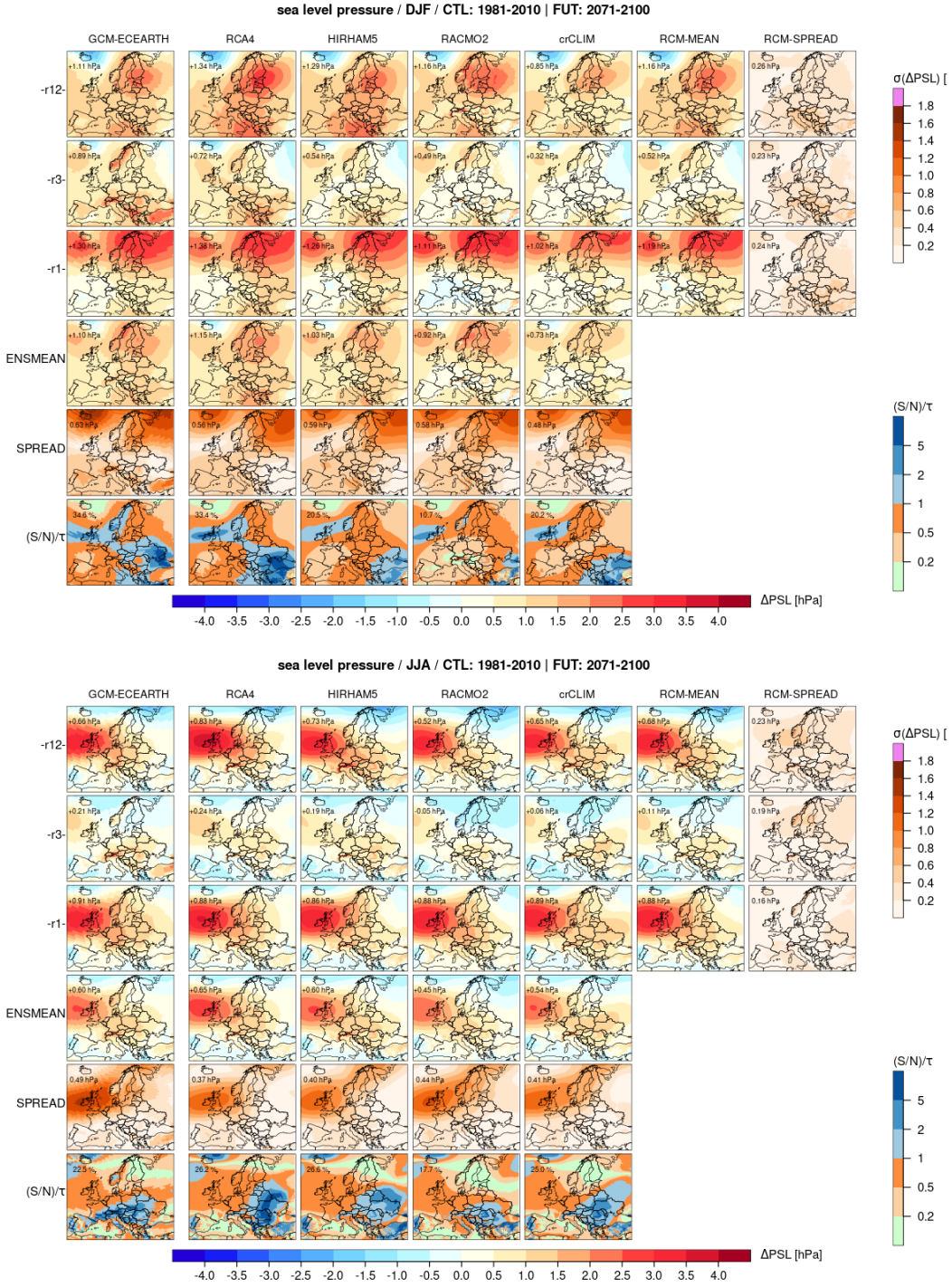

Figure 2. Response in 30-year mean sea level pressure within the M1 driven ensemble for DJF (top figure) and JJA (bottom figure) Columns 2 to 5 show results for the RCMs, columns 6 and 7 the RCM-Mean and inter-RCM spread for each GCM realization, respectively. Row Ensmean and Spread show the inter-member mean and spread. S/N is the signal-to-noise ratio; where |S/N| exceeds τ (bluish colour), the null hypothesis of no change is rejected (see text for meaning of τ). The 1st column is the outcome for the GCM driver apart. The values in the panels refer to the mean over the shown area; for S/N it is the portion of grid cells where the change is significant.

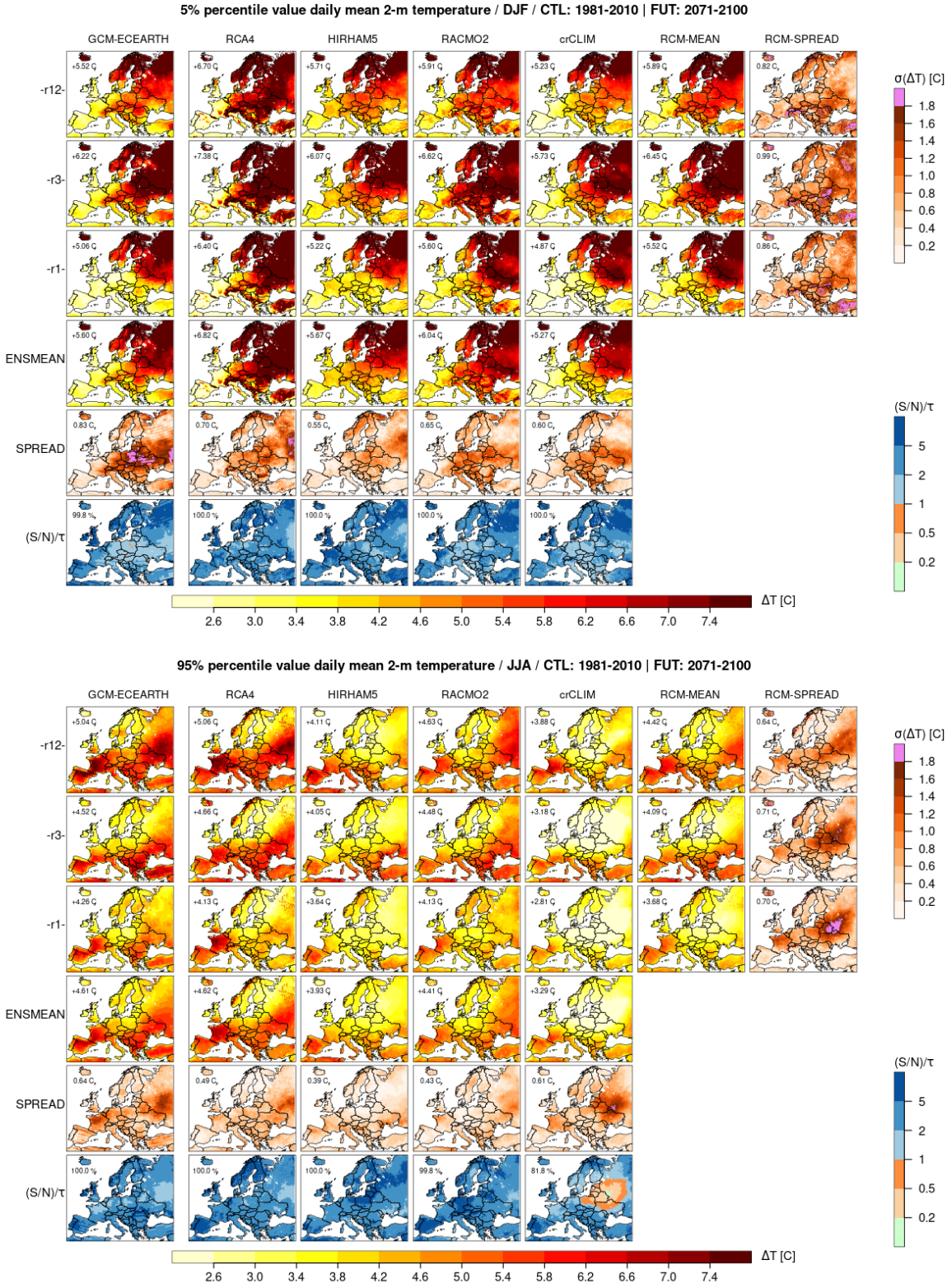

Figure 3. As Figure 2 but for the 30-year coldest day at the P05 level for DJF (top figure) and the 30-year warmest day at the P95-level for JJA (bottom figure), respectively. Numbers apply to the land area.

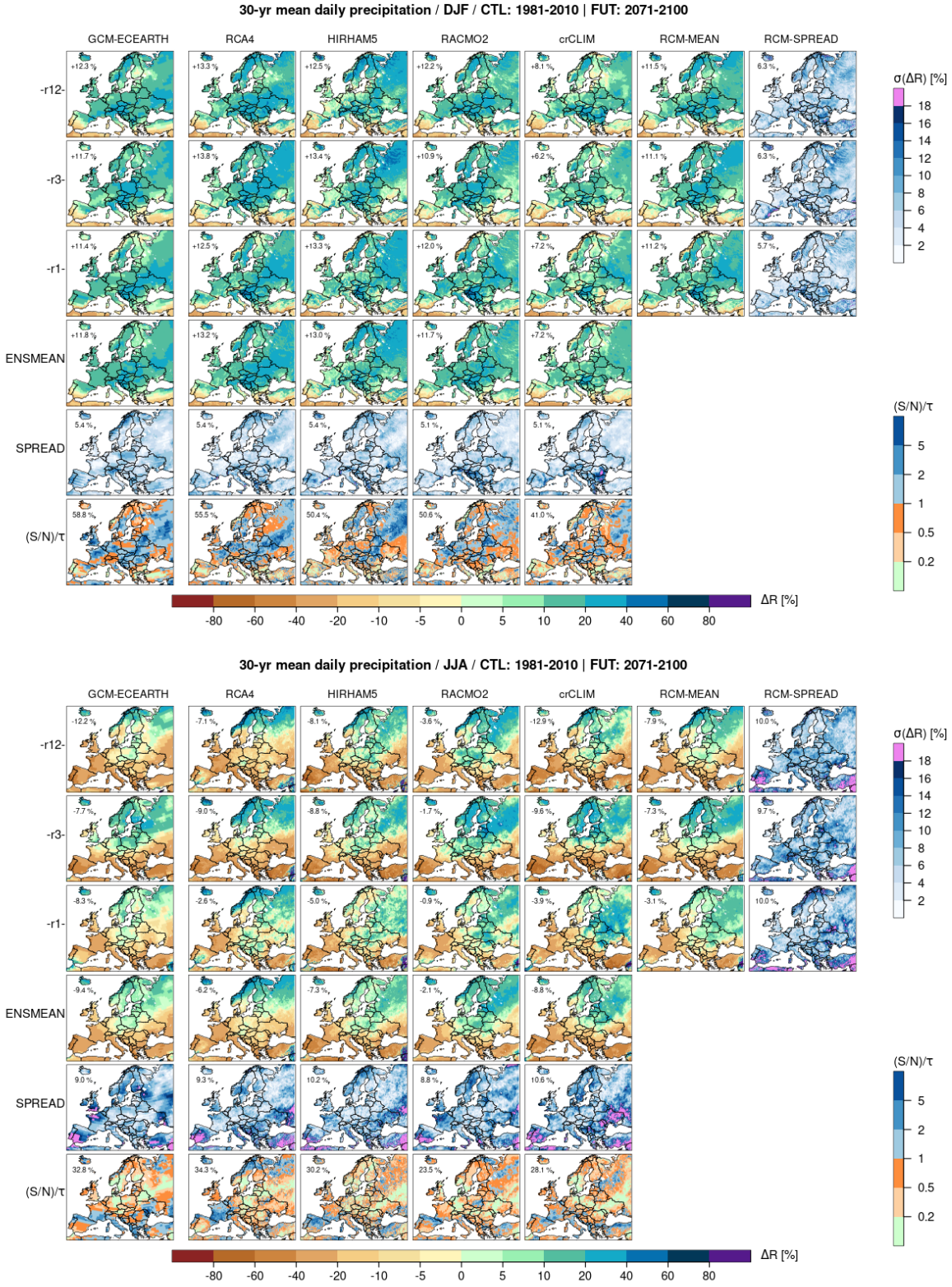

Figure 4. As Figure 2 but for the 30-year daily mean precipitation. Shown is the relative difference in per cent.

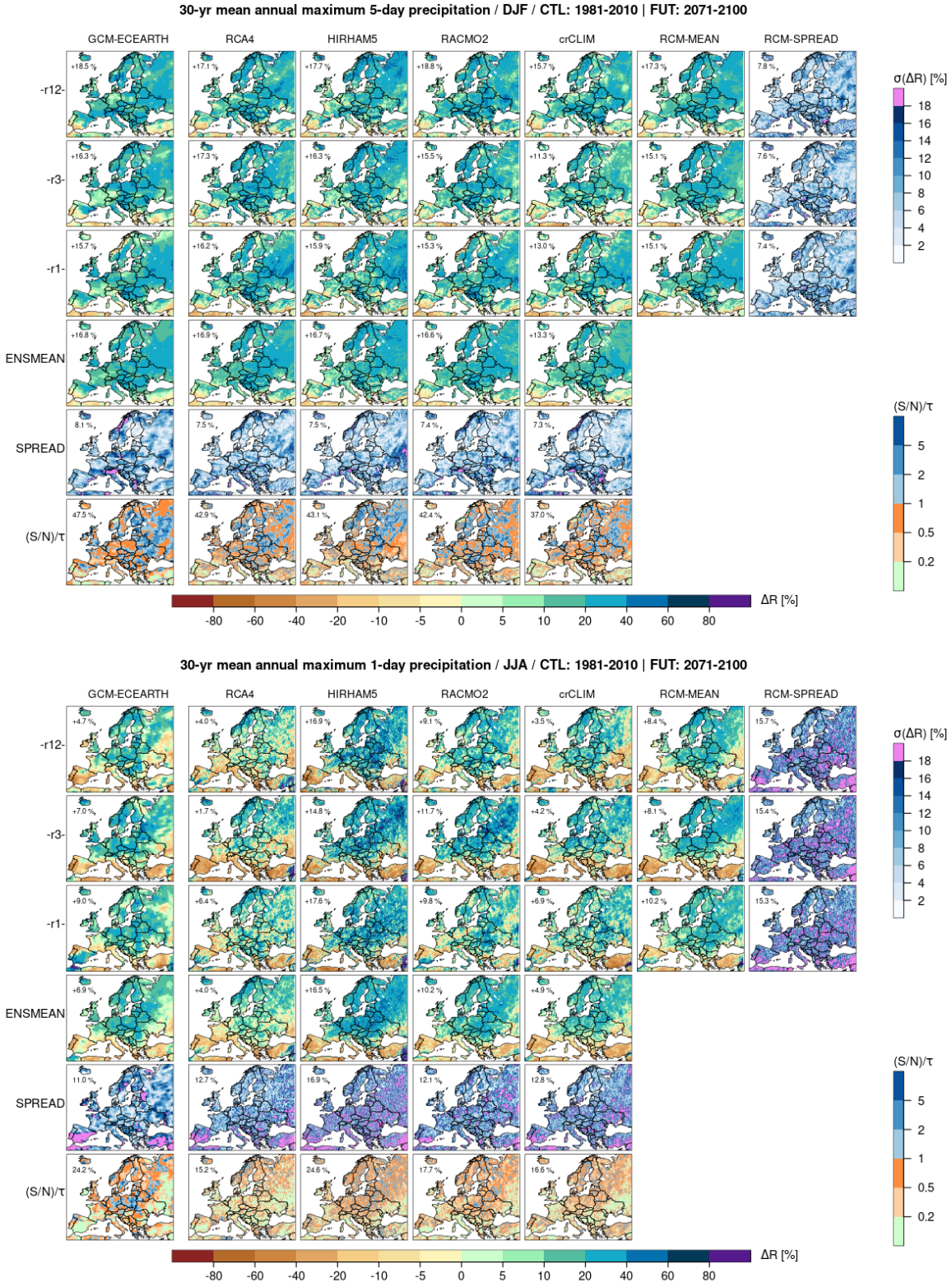

Figure 5. As Figure 4 but for the 30-year mean annual maximum 5-day precipitation amount for DJF (top figure) and the 30-mean annual maximum 1-day precipitation amount for JJA (bottom figure), respectively.

The response in mean precipitation in the M1-driven ensemble for both seasons is shown in Fig. 4. For DJF the response among the RCMs is quite consistent across the realizations, and the inter-RCM spreads are very similar to the inter-member spreads. Yet, the amount of area where S/N is not significant is considerable which is either due to a relatively small signal (Mediterranean), or a large spread (Northern Scandinavia and the Baltic states) or both. For JJA, one RCM (RACMO2) yields a consistently smaller negative response, when spatially averaged, than the other three RCMs; at the same time for some regions (Northern France, British Isles, Southern Spain) the members differ in the sign of the response. Combined this yields again similar sizes of inter-RCM and inter-member spreads, but larger than found for DJF. The ensemble mean response is mostly non-significant except across Southern France and Northern Spain, and in smaller regions scattered across the Mediterranean and Northern Europe. Results for M2 corresponding to mean precipitation are shown in Fig. A3.

More extreme precipitation parameters are shown in Fig. 5, RX5day for DJF and RX1day for JJA (these are the most relevant indicators for hydrological impact studies). The response in maximum 5-day precipitation amount in DJF varies by, on average, an increase of 15 to 19 %, and shows a fairly uniform and spatially smooth distribution across most of Europe. In the ensemble mean, a reduction in RX5day is only seen in the southern parts of the Iberian Peninsula, the north of Spain, and the south of Turkey. In one member (r1) a reduction is also seen in the southern Alpine region and along the west coast of Norway. Like for mean precipitation in DJF the inter-RCM and the inter-member spreads are comparable in size, but sufficiently large to suppress S/N below significance for at least 50% of the land area. Results for M2 are shown in Fig. A4.

The response in RX1day for JJA is by far the noisiest result. In general there is a gradient in response from north-east (high positive) to south-west (neutral to high negative), but spatially the response tends to be patchy, even in the ensemble mean. The difference in RCM signal is substantial with the response in RCA4 and crCLIM relatively small (2-7% increase on average) compared to RACMO2 (9-12%) and HIRHAM5 (15-18%). The inter-member spread for HIRHAM5 is 4-5 percentage points larger than for the other three RCMs. Signal-to-noise in JJA-RX1day for each of the RCMs is significant for not more than 15-25% of the land area, because of the low signal (RCA4, crCLIM) and the large spread (all four RCMs).

4.1.2 Catchment-scale analysis

While Figs 2-5 present domain-wide results at the grid-box level, another approach is to zoom out to larger spatial scales by aggregating detailed grid-box level results over larger areas, while at the same time looking more at details of the temporal evolution of the responses. River catchments provide scales larger than the local (or grid-box) scale, but much smaller than the full regional domain. They also represent natural entities which can be linked to impact hydrology and runoff models. Here we focus entirely on the role of internal variability in determining the response of a selection of temperature and precipitation parameters.

We also compare the outcomes of the seven three-member ensembles (Fig. 1) with results obtained with the 16-member KNMI-EC-Earth-RACMO2 ensemble (Aalbers et al., 2017) and the 50-member CanESM2-CRCM5 ensemble, known as ClimEx (Leduc et al., 2019). The M1, M2, and EC-Earth-driven ensembles are true initial-condition ensembles, in the sense that each of the GCM driven realizations is launched from a coupled state arbitrarily selected from the pre-industrial simulation that had been conducted assuming constant external forcings corresponding to the year 1850. In the ClimEx

ensemble the same procedure was followed to generate a five-member CanESM2-ensemble; subsequently, small atmospheric perturbations were then applied in 1950 to obtain a ten-fold larger ensemble.

Figures 6 and 7 show the development over time of the 30-year running mean change in basin-averaged mean daily precipitation (Rmean) for the winter and summer season, respectively. Individual members are shown in the top panels for each of the seven three-member ensembles built on the M1 and M2 GCM drivers. The lower panel shows the time development of the ensemble properties of the multi-member ensembles. Obviously, there can be large inter-member difference, e.g. during DJF the time paths in all three EC-Earth-r12 driven RCM series strongly deviate from the other two ensemble members (r3 and r1). Interestingly, the deviation is largest at mid-century but disappears at the end of the century. This kind of behaviour is exactly what can be expected in the climate system and what makes such multi-member ensembles so valuable. The variations of these deviations with time are primarily triggered by large-scale internal variability in the climate system as represented by the climate models. For the MPI-ESM-LR members the temporal development is less different, but there is a clear signature coming from the driving member. For the multi-member ensembles the ensemble mean curves are much smoother than any of the curves of individual members, though even the curve obtained for the 16-member KNMI-ensemble is less smooth than one would perhaps have hoped. Yet, the small fluctuations apparent in the catchment-scale response are in line with results from a local-scale analysis (Aalbers et al. 2017) carried out for the same model ensemble. The role of the ensemble size is further illustrated by results obtained by subsampling the multi-member ensembles. Evidently, the mean change derived from 4-member ensembles within the KNMI-ensemble can deviate more than one standard deviation from the full-ensemble mean. And this even holds for 10-member subsamples within the ClimEx ensemble.

For summer (Fig 7) the outcome is somewhat different. On the one hand, the inter-member spreads are smaller resulting in a more robust signal. On the other hand, the impact of the RCMs is larger, in particular in the EC-Earth-driven ensembles, where the RCA4-ensemble shows a much larger drying response than the HIRHAM5-, the crCLIM, and, in particular, the RACMO2-ensemble. The latter is even indecisive on the sign. Here, we remind that the response with less summertime precipitation is least pronounced in RACMO2 among all RCMs (cf. Fig. 4). In this particular region, which is in between areas of strong decrease in the south and increase in the north, the RACMO2 results indicate relatively small changes in precipitation. The reason behind these differences is not analysed in detail here but may involve different treatment of soil-moisture feedback in the different RCMs. For the MPI-ESM-LR-driven ensembles, both inter-member and inter-RCM differences are small. Also the large multi-member ensembles have much less spread in JJA than in DJF. But it invariably holds that for certain periods the estimated mean changes based on subsamples can deviate from the full member response by more than the full member spread. Also, note the large response in the ClimEx ensemble (30% reduction by the end of the century) compared to the KNMI-ensemble (10% reduction) and even to the changes in any of the RCMs and individual members driven by M1 and M2 (at most 20% reduction).

Figures 8 and 9 show results in a similar context as Figs 6 and 7 but for changes in 30-year mean annual maximum five-day precipitation (RX5d) in winter and maximum one-day precipitation (RX1d) in summer. The results for RX5d in DJF (Fig. 8) are qualitatively very similar to those obtained for Rmean. Responses and spreads in both multi-member ensembles have the same characteristics for

both RX5d and Rmean. Differences between RX5d and Rmean are primarily seen in the M1- and M2-ensembles where the role of the RCMs is more discernible for RX5d than it is for Rmean.

The result for RX1d in summer is rather similar to that of RX5d in winter with a positive response in the multi-model ensembles and more often positive than negative changes in the individual members. But like for DJF-RX5d there is a considerable inter-member spread in the multi-member ensembles and there can be substantial positive excursions in individual combinations of RCMs and members which are not seen in the other members, e.g. in M1-r3-HIRHAM5, M1-r3-RACMO, M2-r3-RCA4, and M2-r2-crCLIM. In contrast, reductions up to 10% during multiple decades are seen for M2-r1-RCA4 and, to a lesser extent, M1-r1-RCA4 and M2-r1-REMO. Also worth noticing is that, while the response in JJA-RX1d in the KNMI-ensemble continues to gradually grow towards the end of the century, the response in the ClimEx ensemble starts to decline after 2070, though it is still positive with respect to the reference period at the end of the century. A potential explanation may be that this ensemble ends up with a generally stronger drying than that in the KNMI-ensemble, as indicated by the stronger decline in seasonal mean JJA precipitation (Fig. 7).

Figures 10 and 11 show diagrams of mean precipitation and mean temperature and the (relative) changes therein averaged over the Rhine-Lobith catchment for DJF and JJA, respectively. Clearly, the EC-Earth-driven simulations (orange; red, green, blue, cyan) are always colder than the MPI-ESM-LR (magenta, brown, purple) driven and the ClimEx (steel blue) simulations. RACMO2, to a large extent carrying the same package of physical parameterization as EC-Earth, amplifies this cold behaviour, resulting in profoundly cold simulations for both M1-RACMO2 (blue) and KNMI-RACMO (orange) in either season. The simulations in these combinations are also lowest in precipitation, together with EC-Earth-crCLIM. High values of mean precipitation are found in the M2-RCA4 ensemble in summer and, in particular, in the ClimEx ensemble in winter. Owing to moderate inter-member spreads compared to the inter-model spread the various ensembles emerge rather well distinguishable in the P-T diagram for the reference period (left panel Figs 10 and 11).

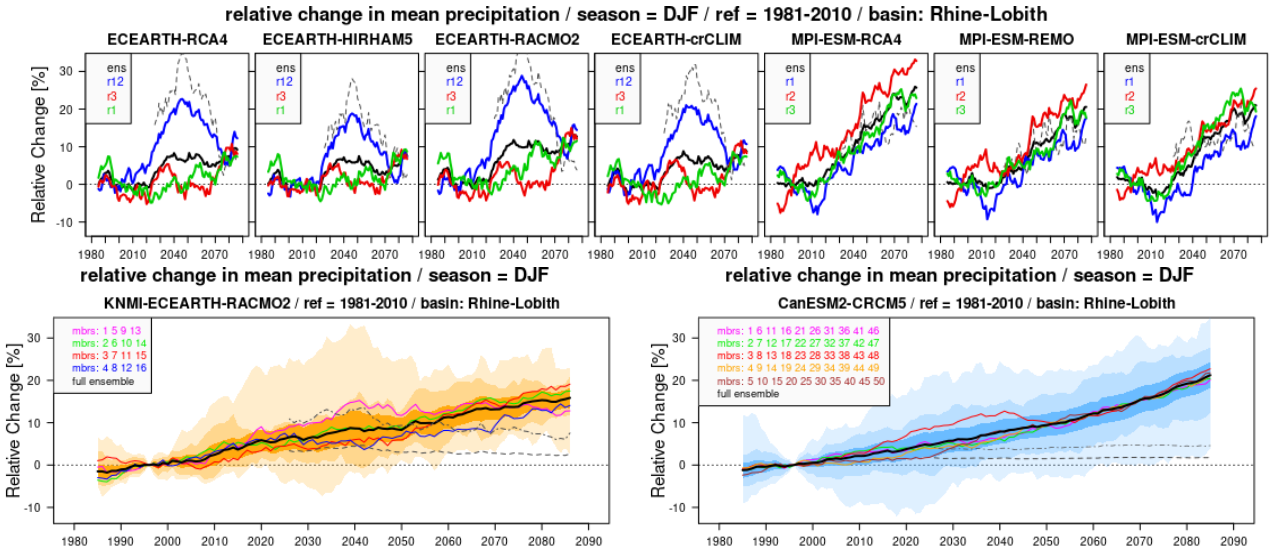

Figure 6. Time series of relative change in 30-year running mean precipitation for DJF with respect to the reference period 1981-2010 averaged over the Rhine catchment upstream from Lobith (sum of German, French and Swiss sub-catchments discharging into the river Rhine). Top row shows the response for the seven M1- and M2-driven three-member ensembles. The black line represents the ensemble mean (S). Bottom-panels show the results for the 16-member KNMI-EC-Earth-RACMO2 ensemble (left) and 50-member CanESM2-CRCM5 ensemble (right). The three levels of shading indicate the standard deviation (N), the range P10-P90, and the full range. The various coloured lines in the bottom panels represent means for multiple 4-member (left) and 10-member (right) subsamples. The dashed curves represent τN with the level of significance corresponding to the number of degrees of freedom. When S exceeds τN the signal is significantly different from zero. The dashed-dotted curves indicate so for the subsamples (only bottom panels).

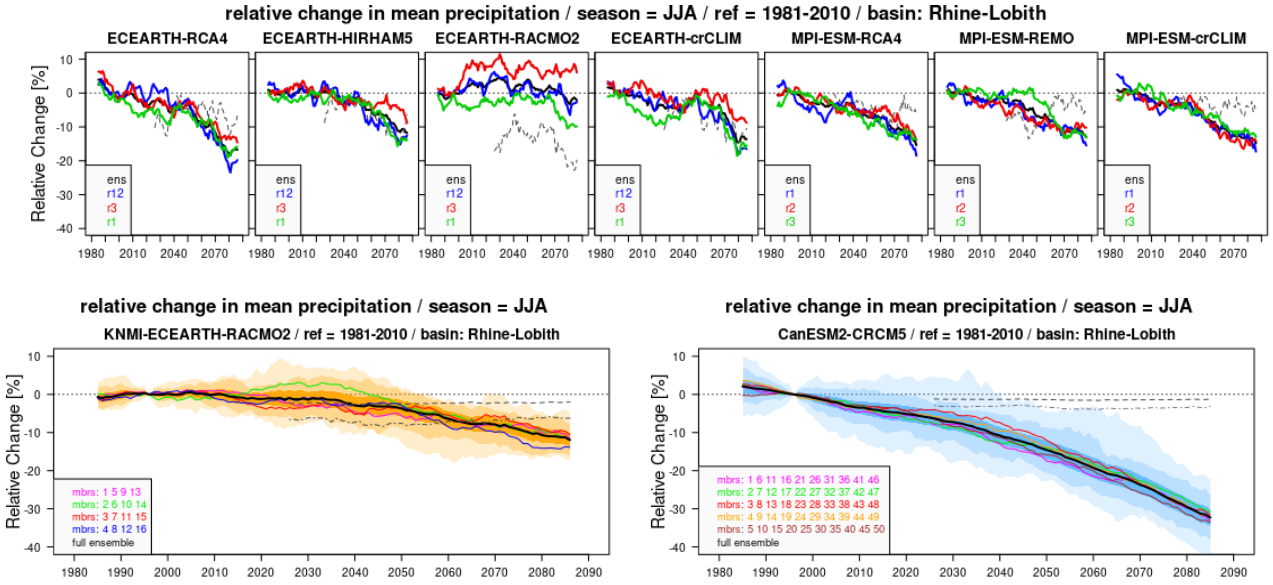

Figure 7. As Figure 6 but for JJA. The dashed curves represent -τN for the full ensemble; dashed-dotted curves for the 4-member resp. 10-member sub-ensembles.

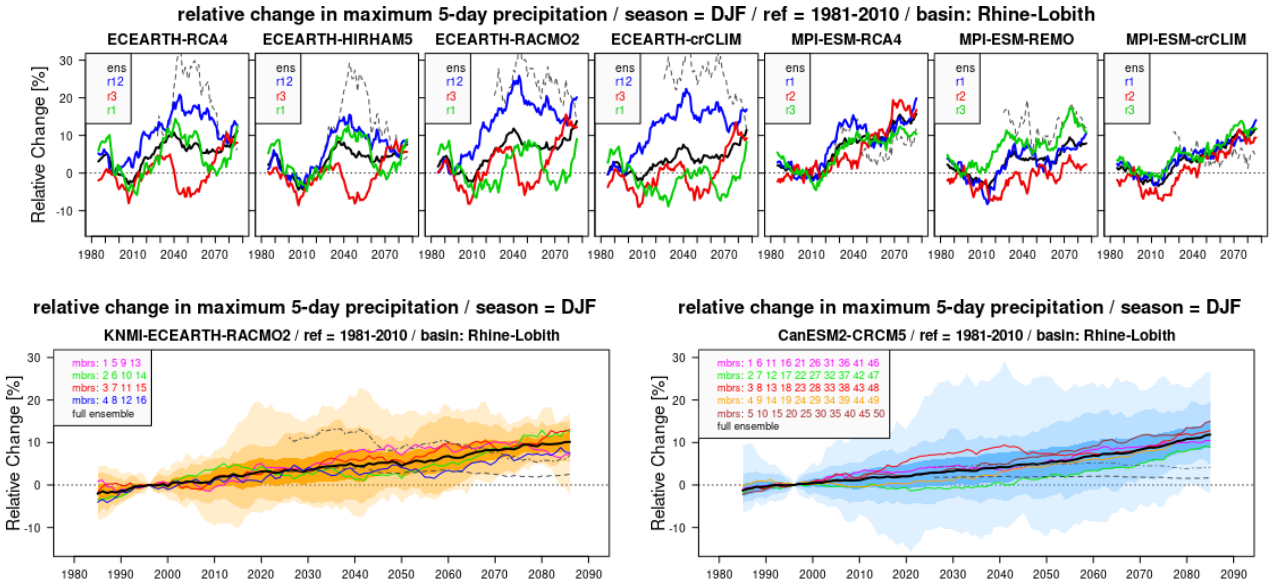

Figure 8. As Figure 6 but for 30-year mean 5-day maximum precipitation for DJF.

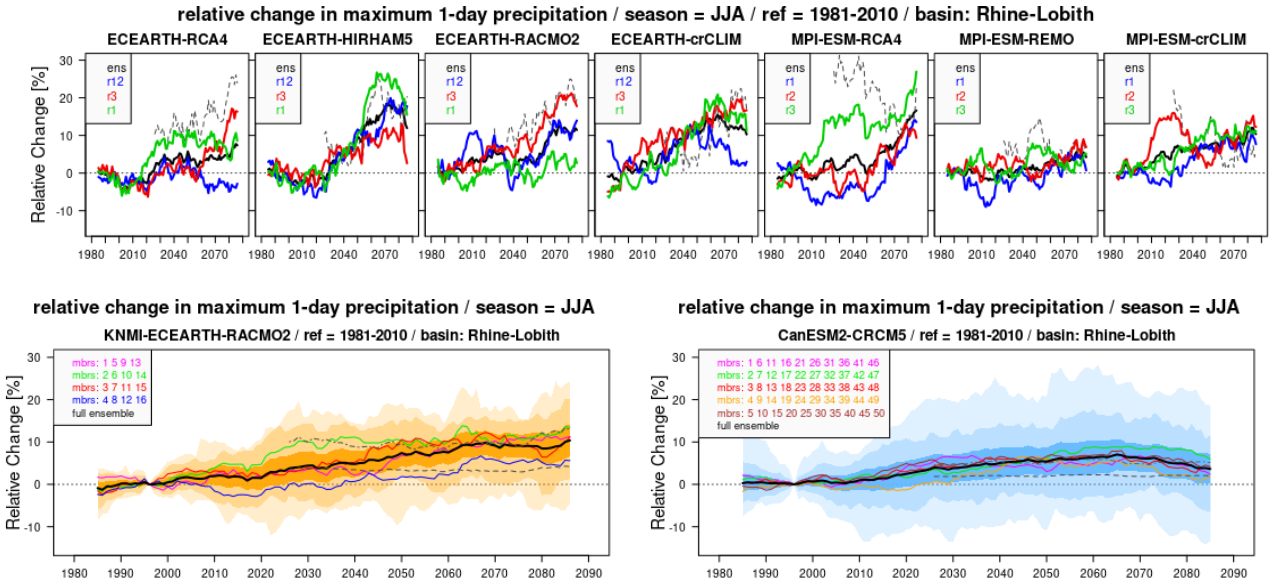

Figure 9. As Figure 6 but for 30-year mean 1-day maximum precipitation for JJA.

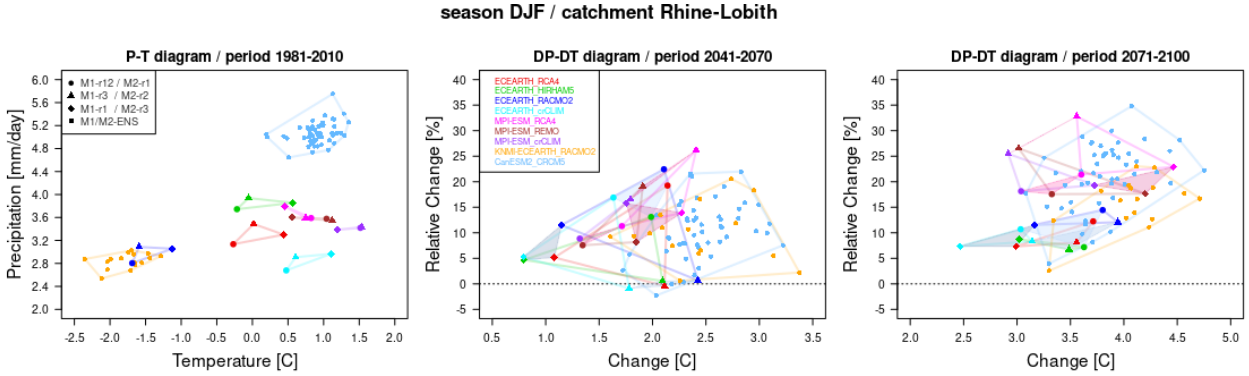

Figure 10. Diagrams of catchment-mean precipitation versus temperature for the DJF season in the reference period (left panel) and their mutual changes for a mid-century (centre panel) and end-of-century period (right panel) compared to the reference period. Symbols depict all individual simulations. Colours represent the different GCM-RCM combinations; symbol types represent the different GCM realizations for M1 and M2. Unfilled polygons envelop the points belonging to the same GCM-RCM combination. Shaded polygons envelop the points belonging to the same GCM-realization. The latter are only shown in the response diagrams for the M1 (grey shading) and the M2 (pink shading) driven ensembles.

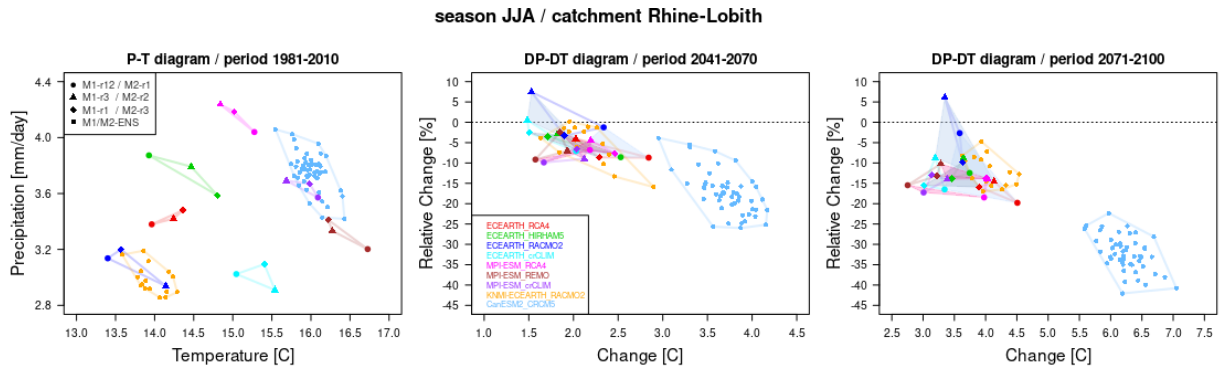

Figure 11. As Figure 10, but for JJA

The change-diagrams are shown for mid-century (2041-2070) and end-of-century (2071-2100; for ClimEx 2070-2099). The ensemble mean and spread in precipitation changes correspond with the time series shown in Figs 6 and 7.

For DJF (Fig 10) the multi-member ensembles (KNMI and ClimEx) and M2-RCA4 show a somewhat larger response in temperature than the other M1- and M2-ensembles, at the same time there is considerable inter-member spread in temperature response. Specifically looking at the M1 and M2-ensembles, the inter-member spread (estimated as the size of the "unfilled" polygons enveloping points with the same colour) for the mid-century period is larger than the inter-RCM spread (estimated as the size of the "shaded" polygons enveloping points with the same symbol type). This applies to both temperature and precipitation and to both ensembles, but to M1 in particular. For the end-of-century period a comparable difference is still found in the M2-ensembles. However, in the M1-ensembles the inter-member spreads are substantially reduced with respect to the mid-century period and have become of similar size as (temperature), or even smaller than (precipitation) the inter-RCM spreads. It is also noted that for the end-of-century period the responses in the M1-ensembles on the one hand and the M2-ensembles on the other have become well separated. Finally, the ensemble mean responses (not shown) in Rmean-DJF range between around 5% and 17% mid-century and 7% and 25% end-century, which is somewhat larger than the inter-member spread averaged over the ensembles. It should be realized that the standard error in the ensemble mean (not shown) is of course much smaller for the multi-member ensembles than for the three-member ensembles as it decreases with the square root of the number of members.

What stands out for JJA (Fig 11) is a clear separation between the ClimEx ensemble and the other ensembles, with very strong warming and drying in ClimEX in comparison to the other ensembles which show much more moderate levels of warming and drying. The different levels of warming can to a large extent be associated with the much higher level of global warming projected by CanESM2 (+4.4 deg. C end-of century) compared to EC-EARTH (+3.2) and MPI-ESM-LR(+3.3). The inter-member spread in JJA-temperature response appears to be somewhat smaller than for DJF. The precipitation response in most ensembles shows a reduction between 10 and 20% with the exception of M1-RACMO2, on the one hand, projecting a reduction of less than 5%, and ClimEx, on the other hand, projecting a reduction of more than 30%. The inter-member spread in JJA-precipitation is smaller than for DJF, with the exception of M1-RACMO2, which is consistent with the result shown in Fig 7. The inter-member spread in the M2-ensembles is smaller than in the M1-ensembles for both temperature and precipitation and in both response periods.

Regarding the M1- and M2 temperature responses for JJA, it is quite interesting to note that in the end-of-century response, the inter-RCM spread (model uncertainty) has increased compared to the mid-century response, and has become larger than the inter-member spread (internal variability) for both the M1- and M2-driven ensembles. This suggests that in a projected much warmer future climate the subsets of RCMs involved in either M1 or M2 have noticeably divergent responses in temperature, potentially owing to considerable differences in the treatment of land-atmosphere coupling (Davin et al., 2016; Vogel et al., 2018). For the response in precipitation we do not find such a clear change in the relative role of inter-RCM and inter-member spread, which may be explained by a larger RCM-induced small-scale variability in summer precipitation compared to temperature (Lucas-Picher et al., 2008; Sieck and Jacob, 2016). Moreover, von Trentini et al. (2020) show that interannual variability changes in response to global warming which will also affect inter-member

spread; however, robustly quantifying the significance of the projected changes in interannual variability requires the use of SMILEs.

As already mentioned in the introduction the use of single-model multi-member simulations offers the possibility to artificially construct very long time series permitting to derive more robust estimates of extreme parameters than feasible for single-member simulations. This is due to the property that all members are produced within the same framework, i.e. an identical model formulation and parameter setting, the same external forcings, and, preferably, the same technical set-up in conducting the simulations. This allows that the inter-member differences can be entirely attributed to internal variability, provided that differences in initial conditions have sufficiently faded away. Without going into full statistical detail Fig 12 illustrates how this works out for the various ensembles regarding extreme basin-mean precipitation, in this case the estimated return time of annual 5-day maximum precipitation across the Rhine-Lobith catchment for a reference and a future 30-year period. The difference in ensemble size is reflected by the range in the time-coordinate which varies from 90 years for 3 members, through 480 years for 16 members to 1500 years for 50 members. While there are differences between the ensembles regarding amount and shape there is a general agreement that the relative signal, i.e. the ratio of future and present-day, varies between 1.05 and 1.10, although for the most extreme events the spread in ratio becomes evidently larger with ratios ranging from less than 1.0 for EC-Earth-HIRHAM5 and CanESM2-CRCM5 to >1.5 for MPI-ESM-RCA4. The exception is the MPI-ESM-REMO ensemble for which the signal is close to 1.0 over the full temporal range. Overall, a qualitative and somewhat crude conclusion might be that the signal becomes robust when the 10 most extreme events are disregarded, which corresponds to a ten times shorter horizon than the full range.

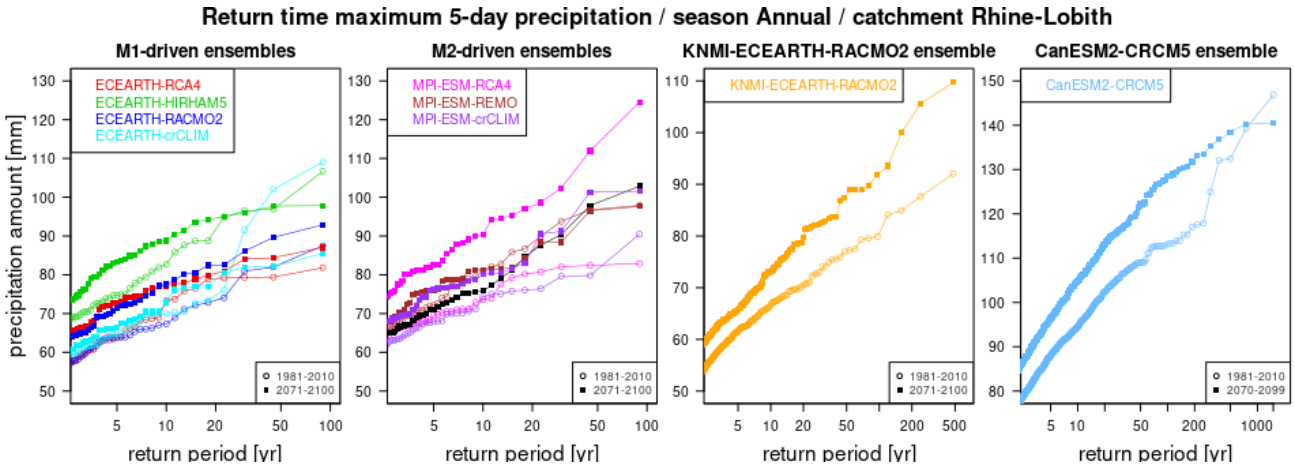

Figure 12. Return time of annual maximum 5-day basin-mean precipitation for the Rhine-Lobith catchment showing the results of the M1- and M2-ensembles (left two panels), and the multi-member ensembles (right two panels) for a future period and the reference period. Note that the ranges along the horizontal axis are different by construction. Also the ranges between the vertical axes are different, but this is done to better see the specific features.

Finally, there appears to be some discrepancy between the results involving RACMO2 (M1 and KNMI), in particular in the responses. This is probably due to several distinct causes. First of all, the domain used in the KNMI-ensemble is substantially smaller than the EUR-11 domain and centred over Western Europe. Secondly, the RACMO2-version used in the KNMI-ensemble employs a slab-ocean

model centred over the North Sea region allowing air temperatures to affect the sea surface temperature (Attema and Lenderink, 2014), whereas the model version used in the M1-simulations adopts the sea surface temperature in the usual way through remapping from the driving GCM member. Finally, the EC-EARTH realizations used for M1 are different from the set used in producing the KNMI-ensemble; in itself this should not be a reason for the perceived difference, but it explains that two members from both ensembles do not have the exact same driver.

4.2 Analysis of Variance

In this section, the two single-GCM matrices will be analysed in an ANOVA framework as in Christensen and Kjellström (2020), i.e., the influence of the period, the GCM choice and the RCM choice will be analysed simultaneously. We will split a field value into contributions from the GCM ensemble member, from the scenario period, and from the RCM as

$$Y_{ijk} = M + S_i + G_j + R_k + SG_{ij} + SR_{ik} + GR_{jk} + SGR_{ijk} \quad \text{(Eq. 1)}$$

where *i* is the period, *j* is the GCM ensemble member, and *k* is the RCM model. This can be done in a unique way given a requirement that all terms sum to zero over each explicit index. So, *M* is the mean over both periods (present and future), all GCM members and all RCMs; *S, G, R* are average effects of period, ensemble member and RCM model. *SG* and *SR* are the difference between future and period average for each member and RCM, respectively; i.e., half the corresponding climate change. *GR* and *SGR* are cross-terms describing the deviation of one simulation from the linear terms for mean climate and climate change, respectively.

Before the analysis, we note that winter temperature in the Barents Sea is very cold for the MPI-ESM-LR r3i1p1 driven simulation with REMO. This simulation is the only REMO-simulation in the ensemble actually performed with the newer REMO2015 model and not REMO2009. While the physical parameterization of the two model versions are so similar that the simulation group at GERICS recommended to view the two models as the same in multi-model inter-comparisons, these results suggest that this assumption should be critically evaluated when studying Arctic winter.

In this report we first discuss the distribution of inter-annual variance for the entire 2x3x4x30 or 2x3x3x30 seasonal mean temperature, precipitation, and wind speed for each period (1981-2010 and 2071-2100 averages are considered in this study), GCM, RCM and year on the various ANOVA terms (Figs 13-15). The impression is that the choice of scenario is dominating over all other terms for temperature, but is much smaller compared to residual (intra-simulation) inter-annual variability for the other variables. Wind speed, compared to precipitation, shows considerably smaller inter-annual variability and larger RCM variability over land; this is explained by the quite different effective roughness length parameters of the five participating RCM models. Both precipitation and wind-speed show considerable RCM variability close to some of the lateral boundaries pointing to different handling of boundary conditions in the models.

Based on these numbers, the formal significance of each ANOVA term (see report C3S_34b_Lot2.1.4.1 and Christensen and Kjellström, 2020) can be calculated and is shown for both matrices and all three fields in Figs 16-21. It should be kept in mind that C3S_34b_Lot2.1.4.1 dealt with several GCM models, whereas the present analysis deals with different ensemble members of one GCM model. The interpretation of the different terms is therefore not the same; the formulae are the same, however. Note also that the significance calculation for each term depends on the number of degrees of

freedom (1 for the linear scenario influence term S; 2 for the global model influence *G,* the regional model influence *R,* the global model influence on climate change *SG,* the regional model influence on climate change *SR*, and 4 for the global-regional model cross-term for mean climate *GR* and the corresponding term for climate change *SGR*); this means that the variability does not have to be responsible for a very large fraction of total variability to be significant.

The two matrices exhibit quite similar mean climate and climate; we shall, however, not be doing any comparisons between the two ensembles, but rather study intra-ensemble properties. We will use the existence of two ensembles to examine features common between the two below.

A typical conclusion from the formal significance analysis of the various terms, common for both matrices, is that the climate change signal is significant everywhere except in areas with very small change, e.g., the zero-line separating negative precipitation signals in the South from positive signals in the North (topmost left panel in Figs. 18 and 19). The cross terms for specific GCM-member-RCM combinations (*GR*) and the corresponding scenario - GCM–member - RCM (*SGR*) terms are basically never formally significant (lowermost two panels on the right for each matrix in Figs. 16-21). The RCM has a formally significant influence on the mean climate (*R*) almost everywhere, but for climate change (*SR*) only for winter temperature over sea and over the central and eastern continent for summer temperature. The influence of GCM ensemble member on climate change (*SG*) is mostly formally insignificant. A notable exception is that the ensemble member choice is significant for temperature change, particularly over the Atlantic Ocean, consistent with the existence of long-term variability exceeding the 30 years used as analysis periods; this long-term variability is also reflected in the pressure change differences of Fig. 2. This happens for both seasons studied. It should be noted that there are several differences between the present analysis and the one in Section 3.1: Here, absolute changes in precipitation and wind are used by necessity due to the formulation of the method. In Section 3.1 the more conventional relative changes are studied. In Section 3.1, only 30-year means are used, whereas the ANOVA analysis includes per-simulation interannual variability in the significance calculation. The latter is probably the reason why several visible differences between ensemble members in Section 3.1 are deemed insignificant in this analysis.

In other words: Climate change in this setting with only one GCM is generally independent on ensemble member and on RCM, with a robust exception for temperature change over sea as well as for summer also over land. The effect of different large-scale ocean and weather conditions between ensemble members does not depend significantly on the choice of regional model, neither for mean climate nor for climate change. One exception is winter mean temperature over the Barents Sea; as explained above, this effect is probably due to differences between two versions of the REMO RCM. The RCM has an important influence on mean climate (see also Figs 10 and 11), but not generally on climate change, except for temperature: in winter over the North Atlantic, and in summer over most of the continent, modulating the relatively large GCM contribution.

Combining the two ensembles with a simple average of significances for the *SG* term (Fig. 22) we see that longer-term oscillations, as manifested in different GCM ensemble members, will have an influence on temperature change in parts of the Atlantic, but not a lot over land. This is especially true for summer. This indicates that decadal circulation variability, as manifested in the differences between GCM ensemble members, is most important over the ocean, and that the effects over land

are larger in winter where weather to a stronger degree is driven by large-scale systems, than in summer where local phenomena are more important.

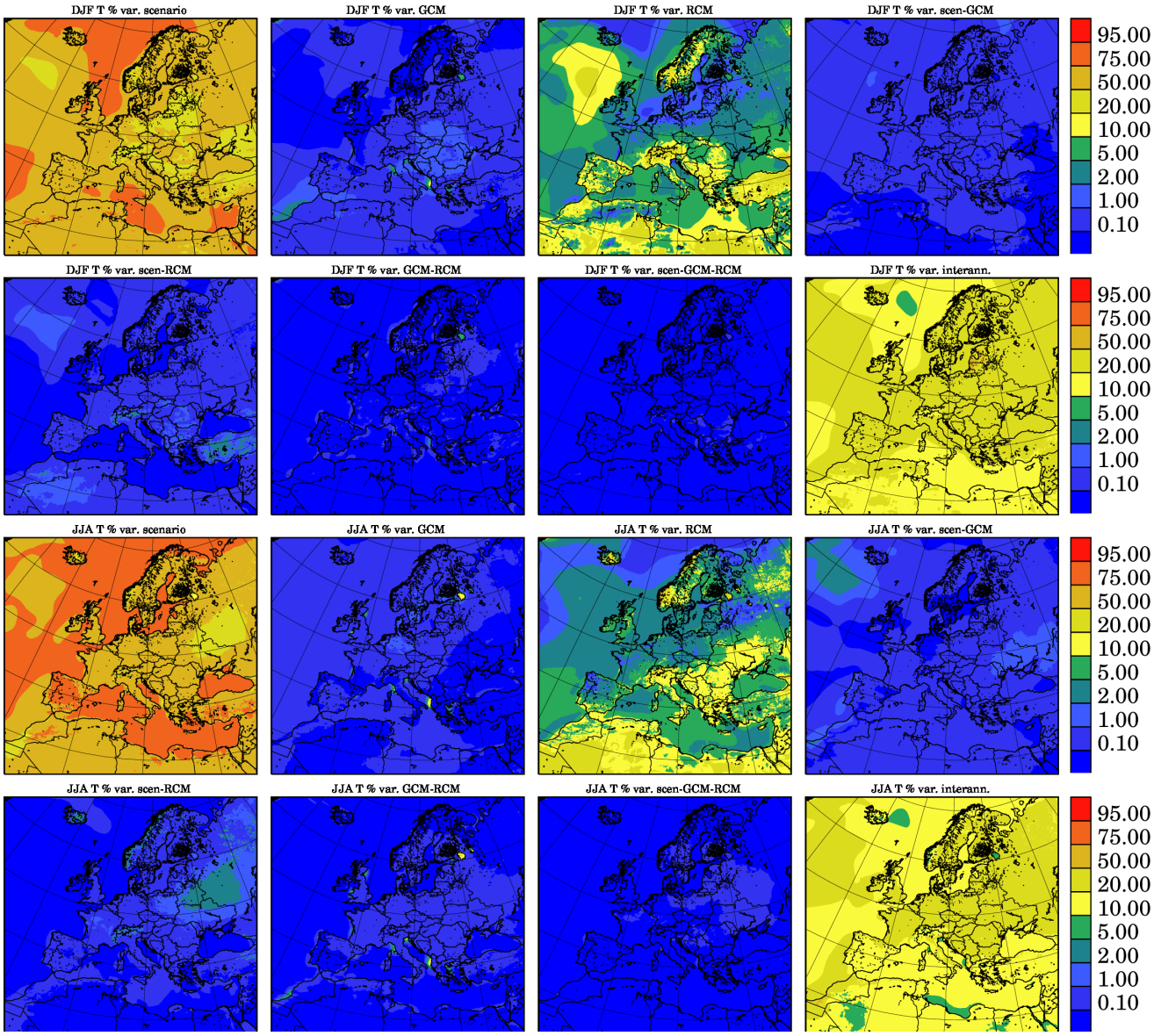

Figure 13 Variability of ANOVA terms as a percentage of total variability for M1. Top two rows: Winter temperature. Bottom two rows: Summer temperature. For each season the terms are: (first and third row from left to right) *S, G, R, SG* and (second and fourth row from left to right) *SR, GR, SGR,* interannual variability. This quantity indicates the relative importance of a term.

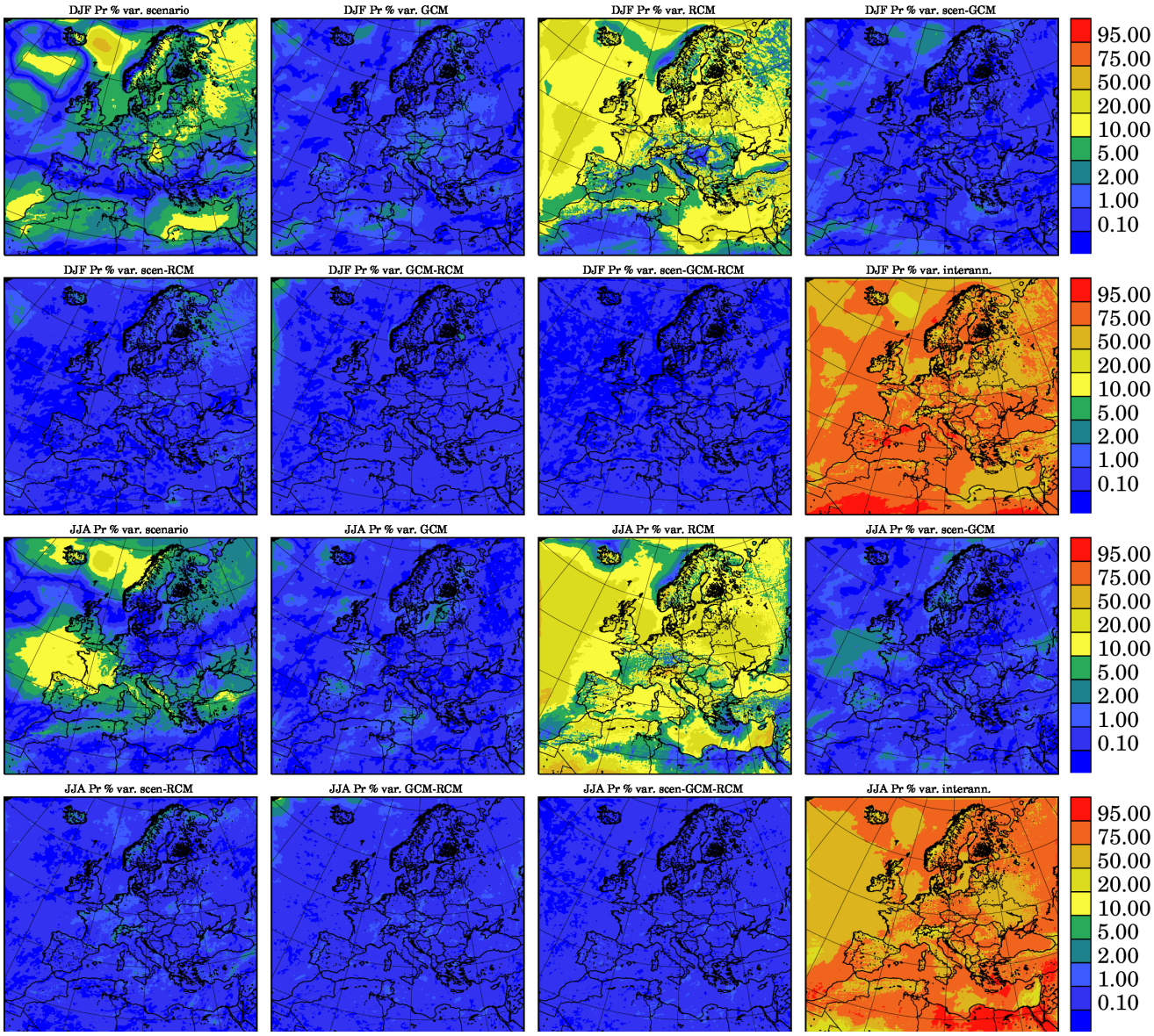

Figure 14 Variability of ANOVA terms as a percentage of total variability for M1. Top two rows: Winter precipitation. Bottom two rows: Summer precipitation. For each season the terms are: (first and third row from left to right) *S, G, R, SG* and (second and fourth row from left to right) *SR, GR, SGR,* interannual variability. This quantity indicates the relative importance of a term.

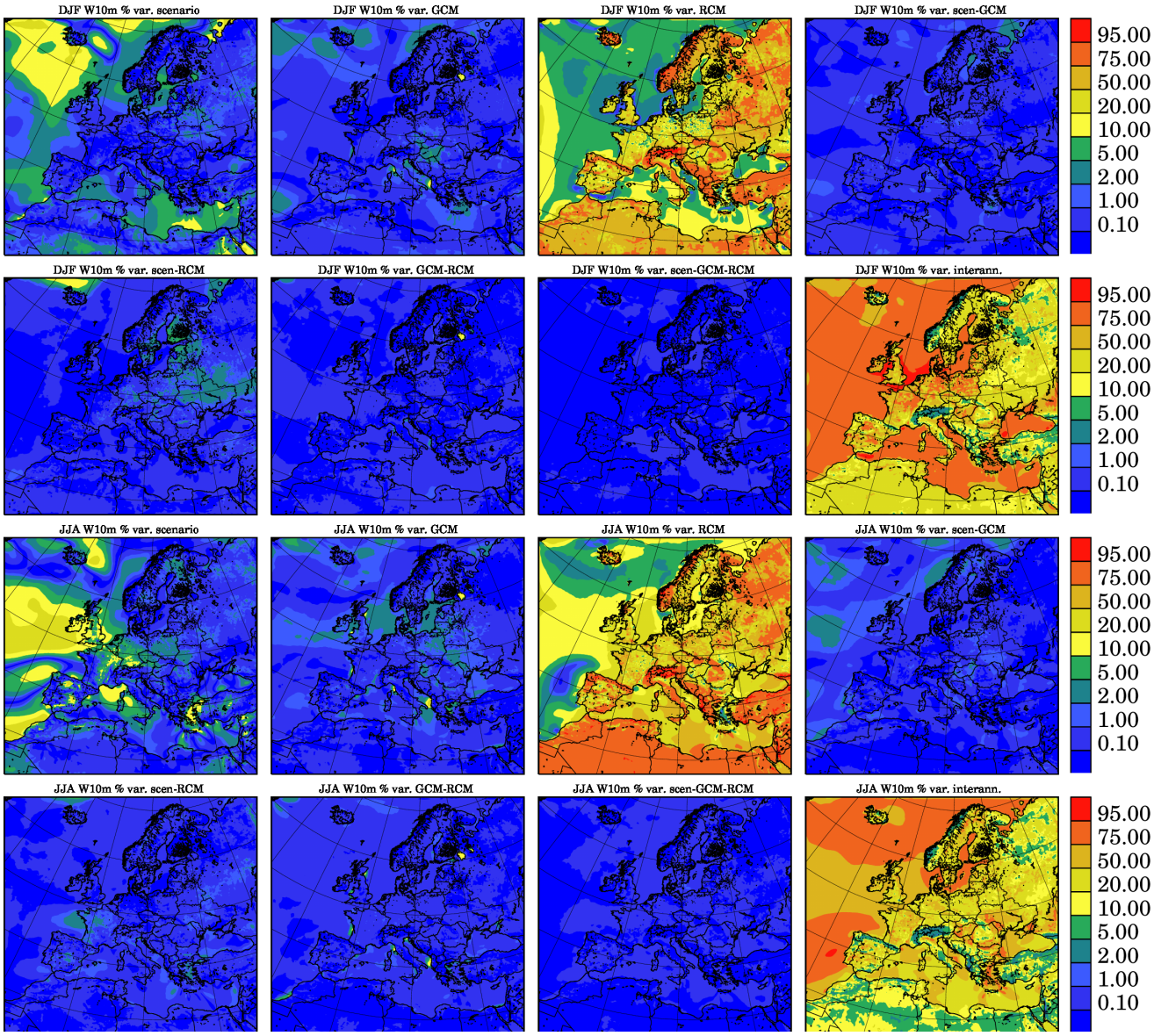

Figure 15 Variability of ANOVA terms as a percentage of total variability for M1. Top two rows: Winter wind speed. Bottom two rows: Summer wind speed. For each season the terms are: (first and third row from left to right) *S, G, R, SG* and (second and fourth row from left to right) *SR, GR, SGR,* interannual variability. This quantity indicates the relative importance of a term.

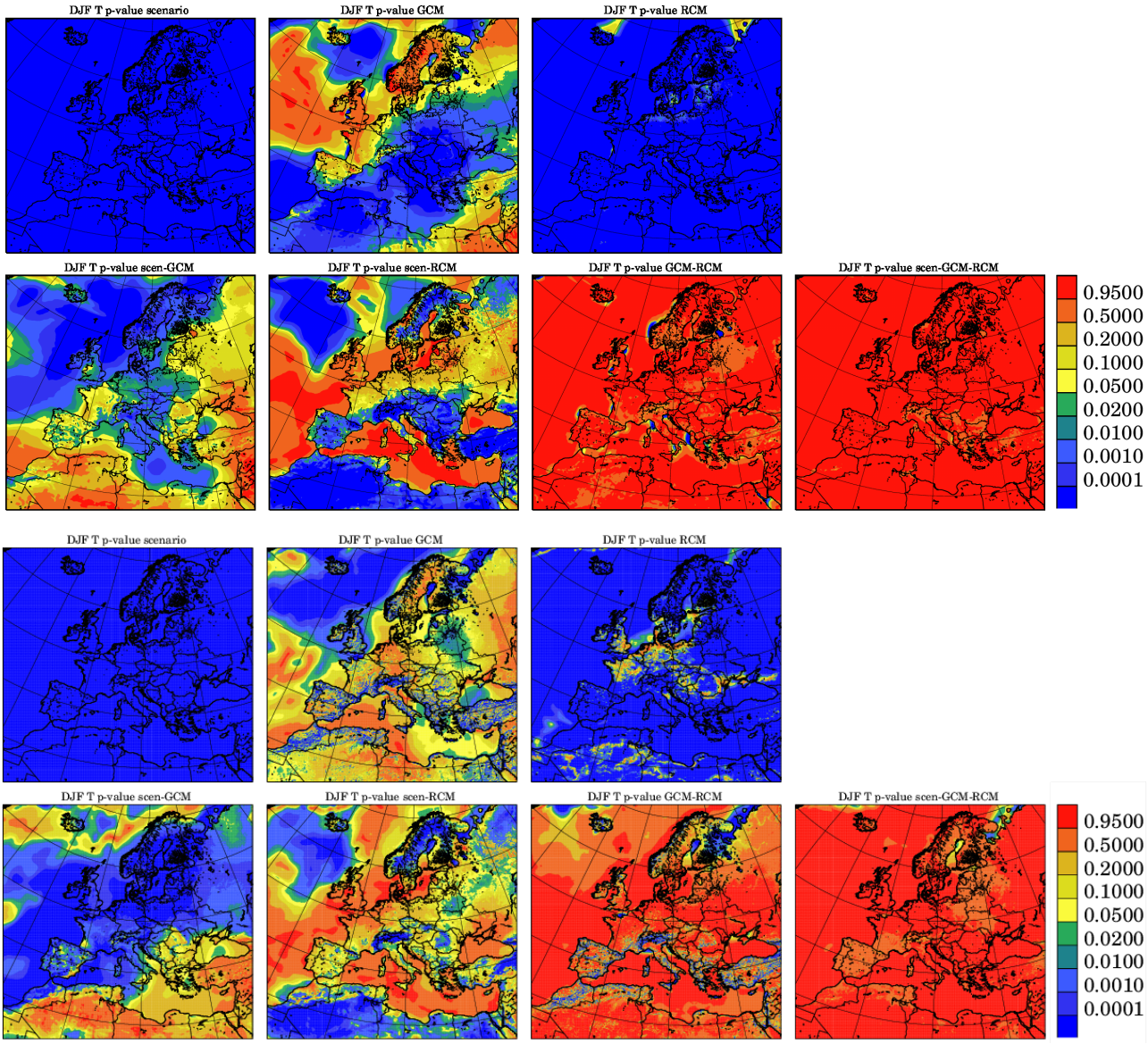

Figure 16 Formal significance of ANOVA terms for winter temperature. Top two rows: M1; bottom two rows: M2. For each GCM ensemble the terms are: (first and third row from left to right) *S, G, R,* and (second and fourth row from left to right) *SG, SR, GR, SGR*. Green and blue colours means the term is formally significant at a 95% level.

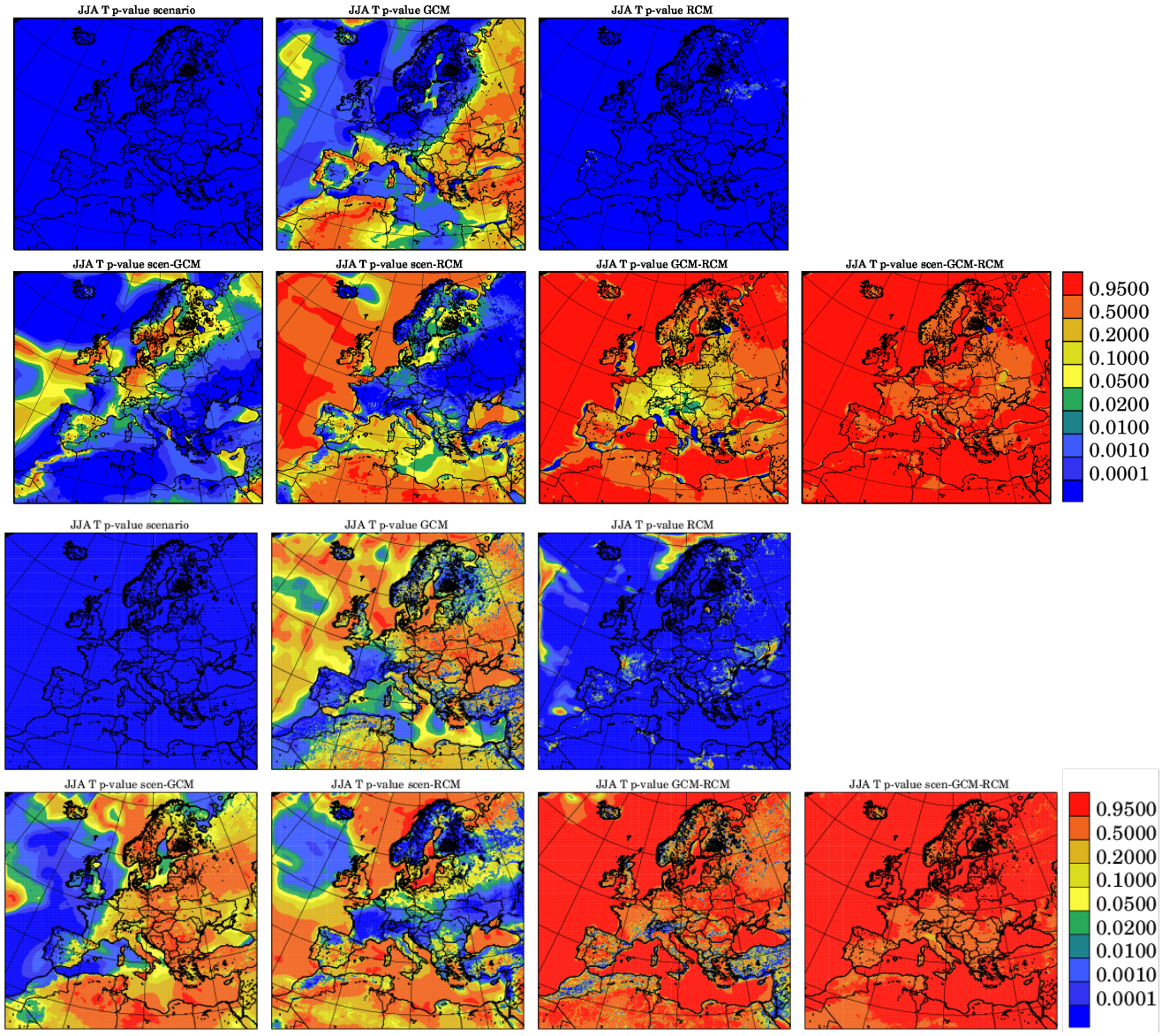

Figure 17 Formal significance of ANOVA terms for summer temperature. Top two rows: M1; bottom two rows: M2. For each GCM ensemble the terms are: (first and third row from left to right) *S, G, R,* and (second and fourth row from left to right) *SG, SR, GR, SGR*. Green and blue colours mean the term is formally significant at a 95% level.

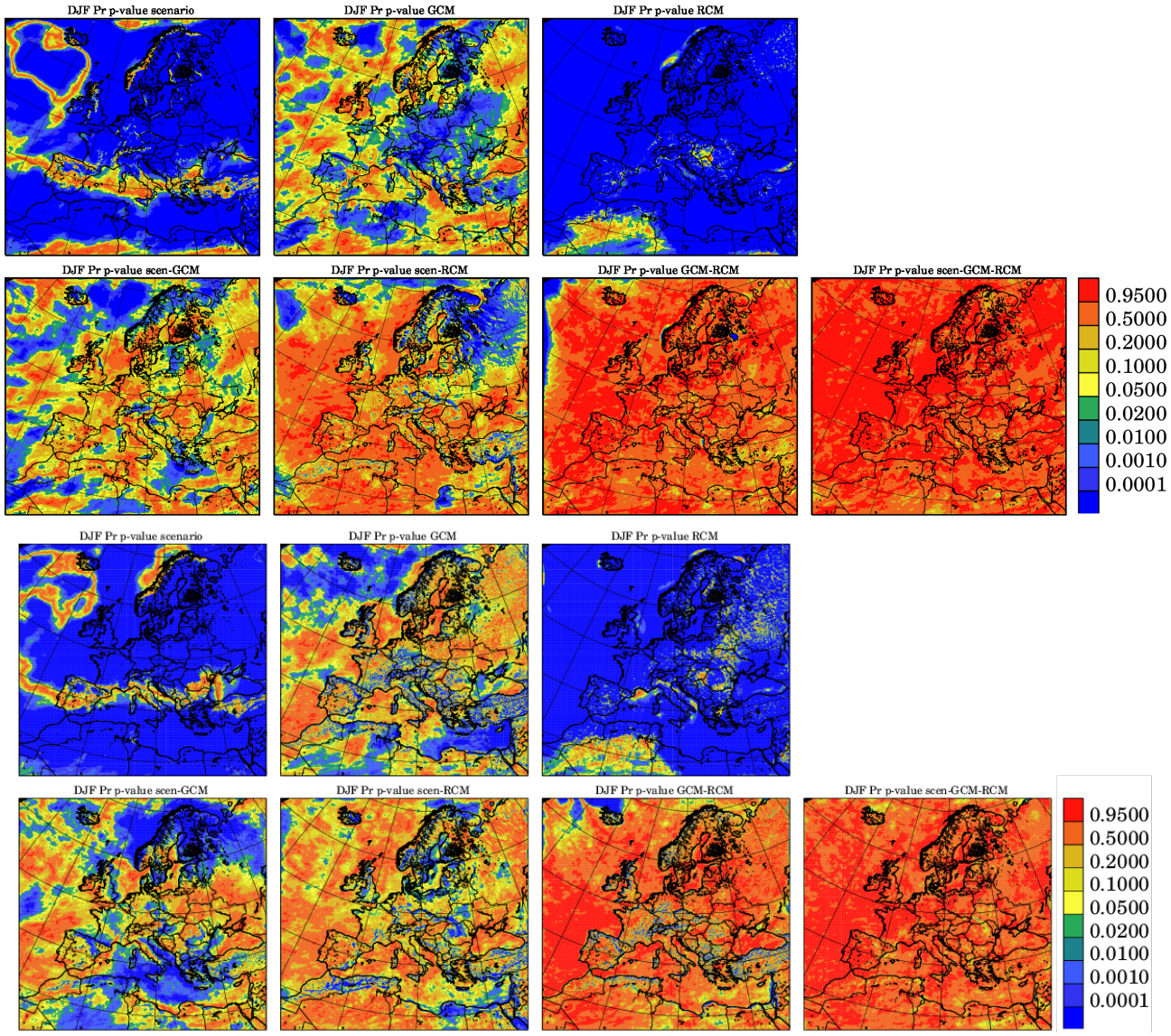

Figure 18 Formal significance of ANOVA terms for winter precipitation. Top two rows: M1; bottom two rows: M2. For each GCM ensemble the terms are: (first and third row from left to right) *S, G, R,* and (second and fourth row from left to right) *SG, SR, GR, SGR*. Green and blue colours means the term is formally significant at a 95% level.

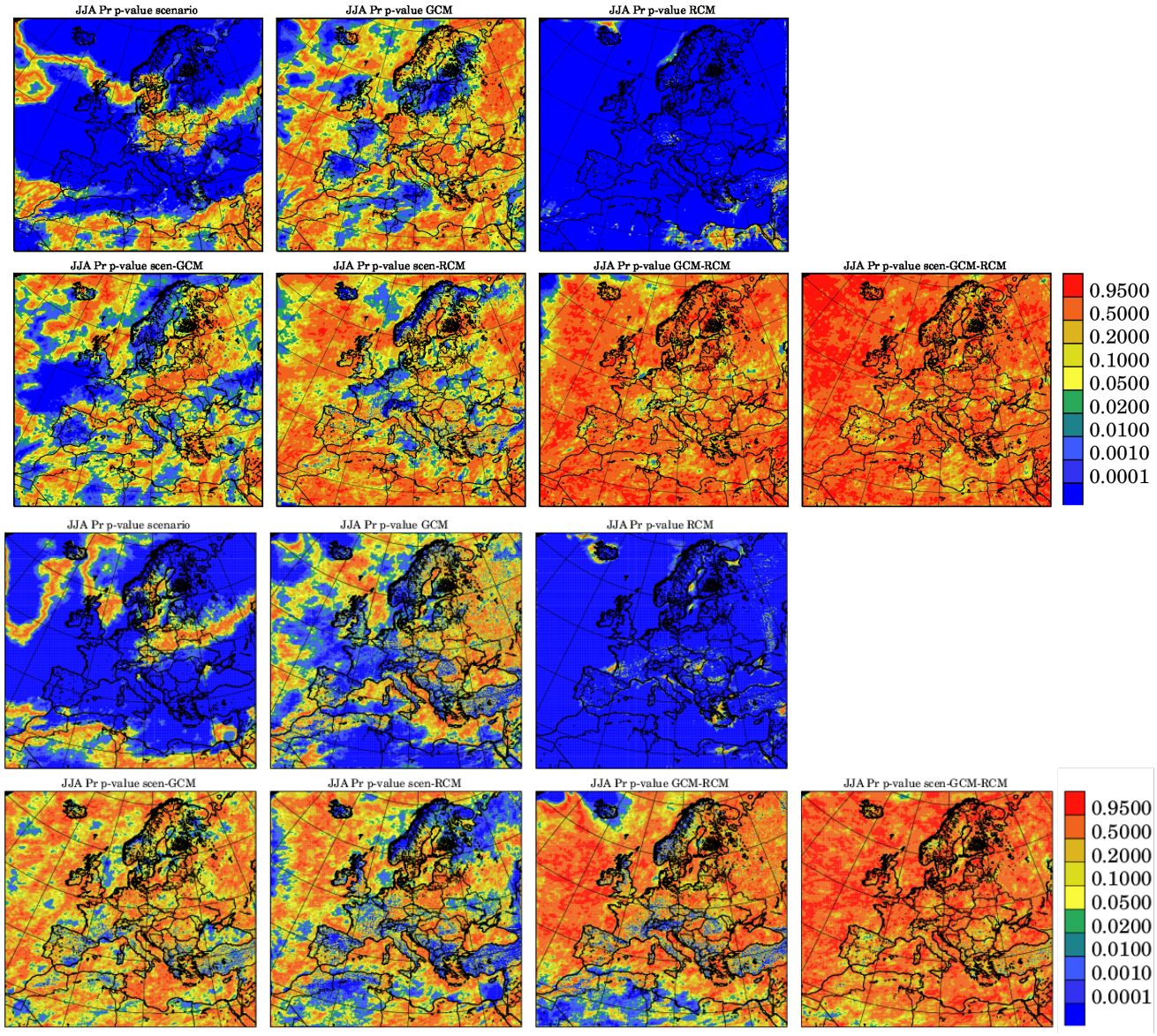

Figure 19 Formal significance of ANOVA terms for summer precipitation. Top two rows: M1; bottom two rows: M2. For each GCM ensemble the terms are: (first and third row from left to right) *S, G, R,* and (second and fourth row from left to right) *SG, SR, GR, SGR*. Green and blue colours means the term is formally significant at a 95% level.

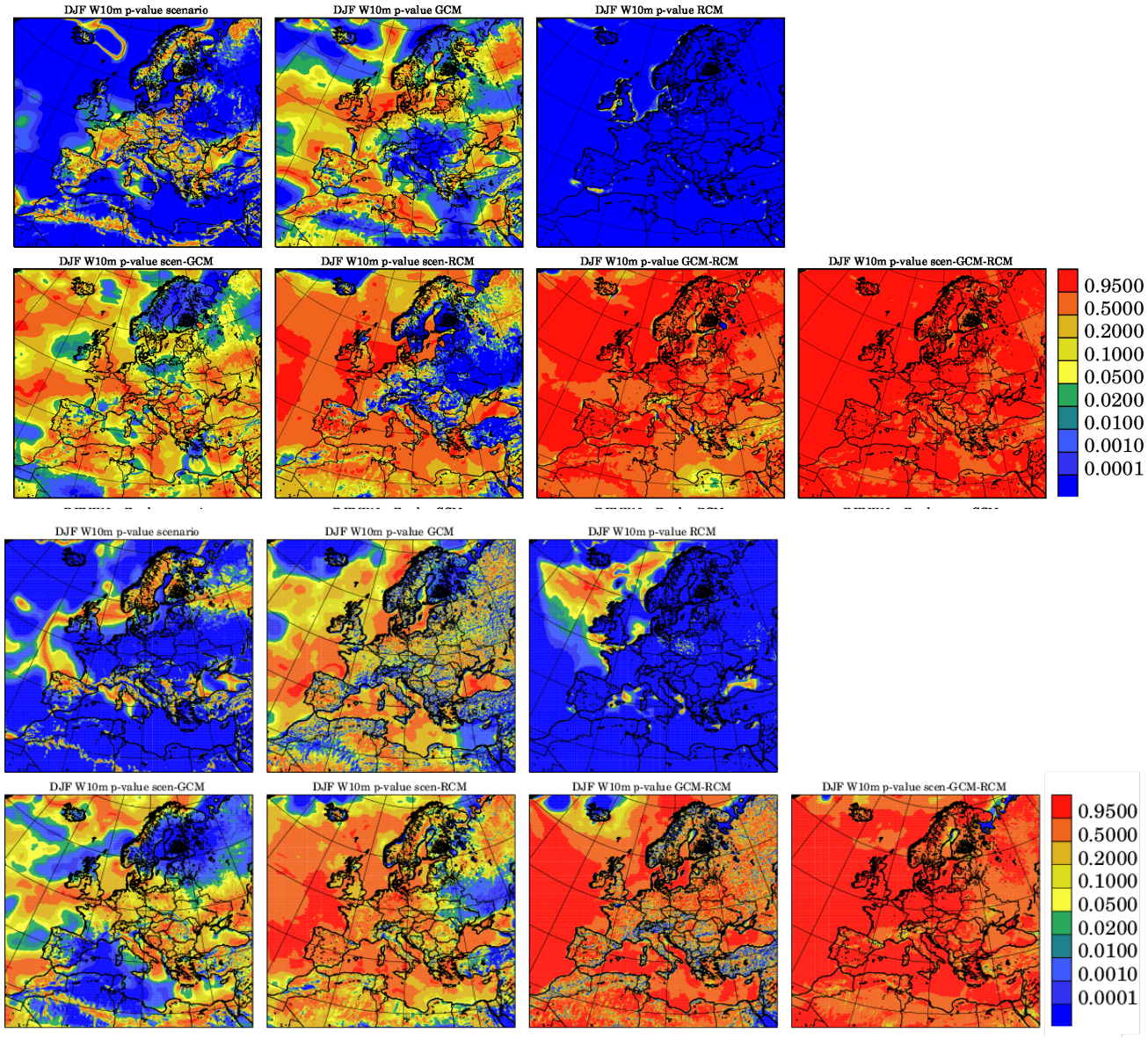

Figure 20 Formal significance of ANOVA terms for winter mean 10m wind speed. Top two rows: M1; bottom two rows: M2. For each GCM ensemble the terms are: (first and third row from left to right) *S, G, R,* and (second and fourth row from left to right) *SG, SR, GR, SGR*. Green and blue colours means the term is formally significant at a 95% level.

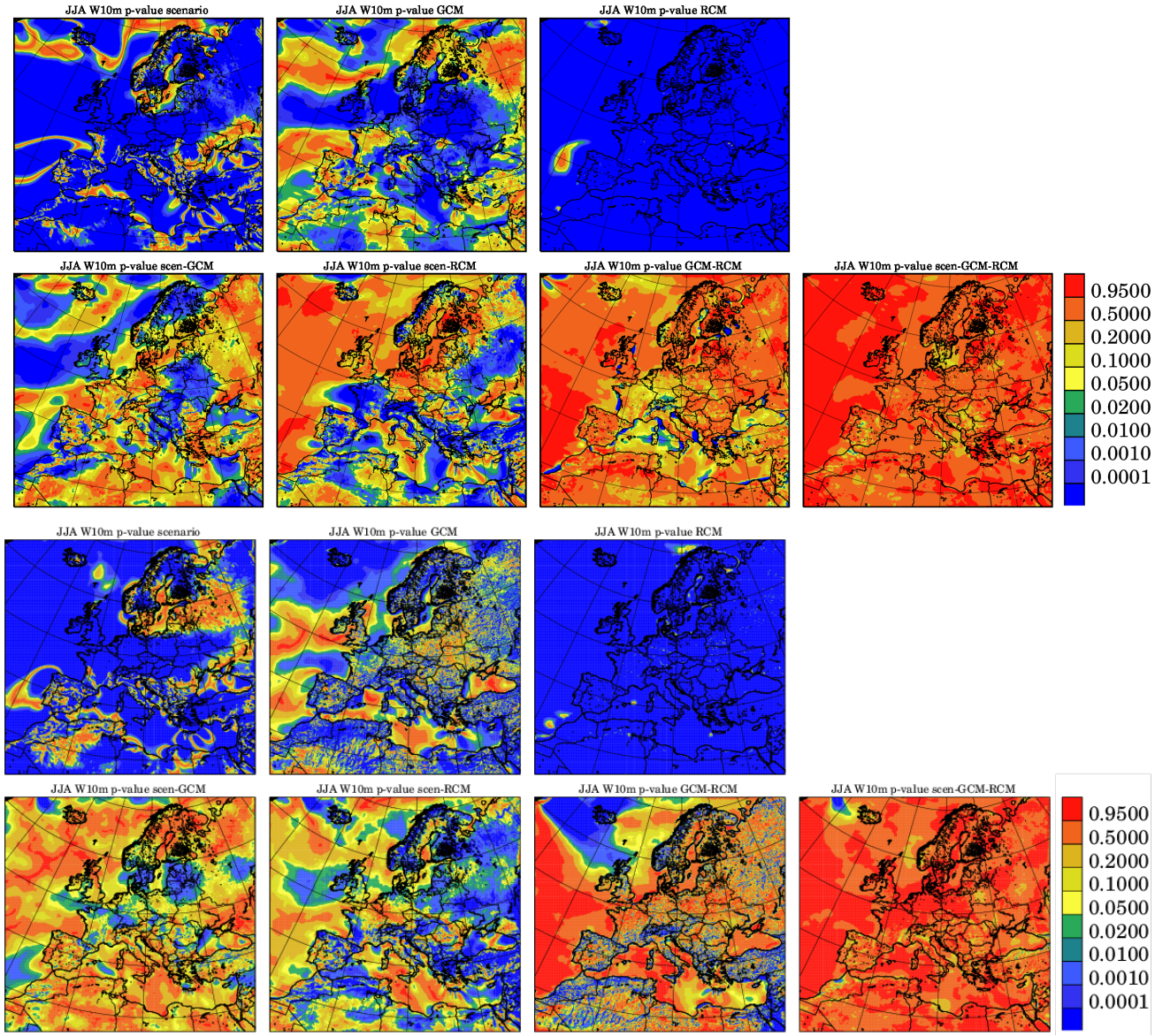

Figure 21 Formal significance of ANOVA terms for summer mean 10m wind speed. Top two rows: M1; bottom two rows: M2. For each GCM ensemble the terms are: (first and third row from left to right) *S, G, R,* and (second and fourth row from left to right) *SG, SR, GR, SGR*. Green and blue colours means the term is formally significant at a 95% level.

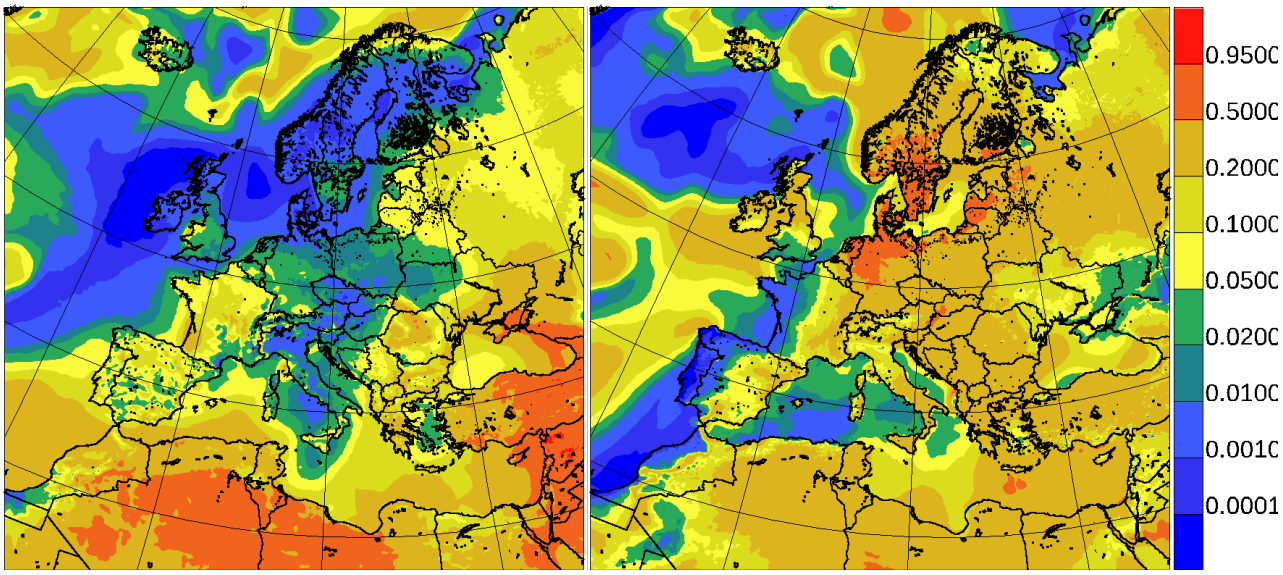

Figure 22 Averaged test probability for the two matrices for the ANOVA temperature SG term. Left: DJF, right: JJA.

5. Summary

Five PRINCIPLES partner institutions have used their respective RCMs to downscale three ensemble members of one or two GCMs generating a total of 21 simulations studied as two separate simulation matrices (3x4 simulations with EC-EARTH (M1) and 3x3 simulations with MPI-ESM-LR (M2); see Fig. 1) with two RCMs, the RCA4 and the crCLIM RCMs, being present in both ensembles. All simulations cover the period 1971-2100; in this study we have compared a control period 1981-2010 and a future scenario (RCP8.5) period 2071-2100.

A full domain grid-scale analysis of climate change responses in both the M1- and M2- driven simulation matrices shows that the **internal variability (inter-member spread) at 30-year time scales in mean sea level pressure, being a measure of large-scale circulation is much larger than the inter-model differences (uncertainties coming from the model formulations)**. This is found for the two driving GCMs, each showing remarkably large inter-member spread in both the summer season (JJA) and, to a larger extent, the winter season (DJF).

The role of circulation variability is reflected by the inter-member spread in the change of high- and low-end percentiles of temperature and in the relative change in mean precipitation, but for these parameters **the model dependence is much stronger than for circulation, resulting in similarly sized inter-model spread**. For extreme precipitation both the long-term internal variability (inter-member spread) and the inter-model spread are proportionally and substantially larger than seen for the other parameters, and, in particular for JJA, large variations are apparent at small spatial scales.

For all three-member ensembles **the resolved-scale signal is nearly everywhere significant for extreme temperature and virtually nowhere significant for extreme precipitation in summer**. For all other parameters it is a mixed picture of alternating regions beyond or below significance. For temperature the high signal-to-noise ratio is associated with the strong external forcing of the adopted high-end scenario; for large-scale circulation the signal-to-noise ratio is almost entirely

determined by the driving GCM-ensemble. These results hold without exception for both sets of three-member ensembles. The inter-ensemble differences in signal-to-noise ratio are relatively small and more related to the driving GCM than to the downscaling RCM.

Within the framework of a catchment-scale analysis the results from the three-member ensembles produced in PRINCIPLES have been compared with results from two single-model multi-member ensembles available from other projects. The analysis focuses on the temporal evolution of climate change in precipitation parameters and the combined change in mean temperature and precipitation. What stand out are the **large inter-member differences in (relative) change between a future and a reference period** owing to long-term fluctuations throughout the full period of simulation. These differences can be related to the role of a driving GCM member (e.g. change in mean winter precipitation in EC-EARTH-r12 forced simulations evolve very differently from the realizations r1 and r3; see Fig. 6) or are seen in specific GCM-RCM combinations. The two larger multi-member ensembles show much smoother response curves than the 3-member ensembles, yet **even the 16-member ensemble mean response is not free of small but discernible fluctuations**. In fact, **climate change signals derived from subsamples can deviate from the full ensemble climate change signal by more than one standard deviation during prolonged time intervals**; this even holds for 10-member subsamples selected from the 50-member ensemble. This shows **the importance of using multi-member ensembles** for an accurate estimation of the climate change response, even at the scale of a large river catchment.

The inter-member spread is smallest in relative change in mean precipitation in summer in all ensembles; the spread in relative change in mean precipitation in winter and in extreme precipitation in both seasons is larger and similar in size. All ensembles show, averaged over the Rhine-catchment, a decrease in mean precipitation in summer and an increase in mean precipitation in winter. In extreme precipitation all ensembles project an increase in both seasons, but the response in maximum 1-day precipitation in summer in the ClimEx-ensemble shows a clear transition from an increase to a decline during the second half of the century, though at the end of the century the response is still positive.

Regarding the combination of temperature and precipitation, the reference state shows considerable inter-model differences. All model ensembles driven by EC-EARTH are low in temperature and relatively dry; for the combinations including RACMO2 this behaviour is amplified. The ClimEx-ensemble, on the other hand, is higher in temperature and relatively wet; winter precipitation, in particular, is very high. The MPI-ESM driven ensembles fall somewhere in-between. Because the inter-model differences are larger than the inter-member differences for the mean climate, the various ensembles are generally well separated in the P-T diagram. In the future combined change (DP-DT) diagrams internal variability (inter-member spread) has a more prominent role, but still the responses of the different ensembles tend to fall in distinct point clouds. This is clearest in summer and in the end-of-the-century period. In fact, the ClimEx point cloud in summer becomes completely disjunct from the other ensembles. Finally, in **both the EC-Earth and the MPI-ESM driven three-member ensembles the inter-RCM spread in summer temperature response for the end-of-the-century period is found larger than the inter-member spreads, implying that uncertainty due to model differences has become larger than uncertainty due to internal variability**. A potential explanation is that land-atmosphere coupling in the various RCMs is responding quite differently under future climate warm conditions.

In conclusion, the analysis of a series of RCM ensembles produced within PRINCIPLES is compared with results obtained from two multi-member ensembles. Obviously, large ensembles of single-model systems provide the cleanest way for estimating internal variability across a range of temporal scales. But this requires a large effort in terms of computation and data storage, which is only feasible in dedicated studies. What we can learn from this analysis is that, **while there can be distinct inter-model differences in the representation of the climate and climate change, the size of the inter-member spread appears to be rather insensitive to the model specifics, and much more determined by the climate parameter that is examined**. **This indicates that downscaling a selection of GCM-SMILEs with a single RCM or a few RCMs may be enough to yield an accurate estimate of the inter-member spread that is representative for RCMs in general**. Alternatively, the potential amount of inter-member spread might be estimated from the inter-annual variability (Aalbers et al., 2017), but this can only provide useful results when scales of variability are shorter than decades.

The ANOVA analysis of the results share many conclusions with the above analysis. Splitting all variability into internal, scenario period, GCM, RCM, and cross-term contributions shows that the **GCM ensemble member has a significant effect on the mean state of the climate for a 30-year time slice for large parts of the integration area for wind speed and precipitation and especially for temperature**. There is, however, very little agreement between the location of these areas between M1 and M2; we therefore conclude that **an ensemble size of 3 GCM simulations is too small to get a good signal-to-noise ratio for the mean climate**. An alternative explanation would be that the geographical distribution of long-term variability is fundamentally different between the two GCM models behind M1 and M2, respectively; this does seem improbable due to the ability of both models to simulate realistic climate variability, but would have to be excluded through analyses of larger GCM ensembles.

As expected (Christensen and Kjellström, 2020) **the mean climate has a robust dependence on the RCM almost everywhere for all fields**. Climate change for averaged climate parameters is significant for most areas, fields, and seasons, with the obvious exception of places where change is very small, e.g. in the transition area between positive and negative precipitation change.

Regarding climate change it is also temperature, which shows the largest effect of long-term fluctuations as reflected in quite large areas of significant differences between ensemble members of the same GCM. The RCM for the most part is formally insignificant for climate change, as opposed to mean climate. **The RCM model has, however, a large influence over parts of the North Atlantic for winter temperature change as well as over most of the continent for summer temperature change.**

There are rather few fields and seasons where the choice of GCM ensemble member plays a significant role for results when their roles in each matrix are compared. The winter temperature *G* term, the ensemble member influence on mean climate is similar for M1 and M2, and we therefore conclude that there will be decadal variability influencing the North Atlantic winter temperature significantly. In contrast, the same term for summer temperature shows large significant areas for both matrices, but the spatial patterns are not the same. Therefore, we conclude that the significance in this case is spurious.

The significance discussed in Section 3.1 corresponds to agreement between ensemble members about climate change, for all of the GCM-RCM three-member ensembles. This corresponds in Section 3.2 to the relation between the *SG* terms in the ANOVA-based analysis. We can compare Fig. 4 to the M1 panels of Figs 18 and 19, as well as Fig. 14, for winter and summer precipitation. Several differences in the types of analyses make it difficult to make a direct comparison. However, the common result for the two analyses of these particular fields agree: **In winter the areas with the largest variability among GCM-ensemble members are located in south-eastern Europe and in northern Scandinavia, but it is smaller than interannual variability** (Fig. 14) and hence only marginally significant in the ANOVA analysis; **in summer, France and Spain as well as parts of Scandinavia and south-eastern Europe show the largest ensemble-member dependence of climate change**, in this case sufficient to be formally significant in Fig. 19.

In broad terms, each GCM ensemble gives internally consistent results about climate change. It is, however, clear from the analysis in Section 3.1 that this is not the case for extreme variables, when regional and local changes are studied, or when temporal trajectories are compared.

These results indicate that **it would be relevant in the design of a multi-model ensemble to also look at some single-GCM sub-ensembles as part of the multi-model ensemble, particularly in the case of extremes**; our results presented here, particularly the frequently large differences between M1 and M2 with respect to the geographical distribution of ANOVA-based significance, indicate that **a three-member ensemble is too small to generate credible maps of inter-member variability**. It should in most cases be adequate to generate such an ensemble with a single representative RCM, as **our results do not show a large cross-dependence between GCM internal variability and the choice of RCM**; however, we have described exceptions from this general statement above, e.g. summer average temperature, and it should also be kept in mind that the RCM itself also has a large influence on the mean climate. In such an analysis it would be important that those RCMs not performing SMILE ensembles will all downscale at least one common GCM ensemble member to facilitate variability analysis.

6. Acknowledgements

The authors would like to thank Ralf Ludwig, Fabian von Trentini and Raul Wood from the Ludwig-Maximilian University of Munich for making available results from the 50-member ClimEx ensemble produced for the European domain with the GCM-RCM combination CanESM2-CRCM5.

7. References

Aalbers, E. E., G. Lenderink, E. van Meijgaard and B. J. J. M. van den Hurk (2017), Local-scale changes in mean and heavy precipitation in Western Europe, climate change or internal variability? Clim. Dyn. DOI 10.1007/s00382-017-3901-9

Attema, J.J. and G. Lenderink (2014), The influence of the North Sea on coastal precipitation in the Netherlands in the present-day and future climate. Clim. Dyn., **42**, 1, 505-519, doi:10.1007/s00382-013-1665-4.

Addor, N., and E. M. Fischer (2015), The influence of natural variability and interpolation errors on bias characterization in RCM simulations, J. Geophys. Res. Atmos., 120, 10,180–10,195, doi:10.1002/2014JD022824

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

Davin, E.L., E Maisonnave, S. I. Seneviratne (2016), Environ. Res. Lett., **11**, doi:10.1088/1748-9326/11/7/074027

Deser, C., Phillips, A., Bourdette, V., and Teng, H. (2012), Uncertainty in climate change projections: the role of internal variability. Clim Dyn 38: 527-546. doi:10.1007/s00382-010-0977-x

Hurk, B.J.J.M. van den, E. van Meijgaard, P. de Valk, K.J. van Heringen and J. Gooijer, 2015: Analysis of a compounding surge and precipitation event in the Netherlands Environmental Research Letters, **10**, 35001-35009, doi:10.1088/1748-9326/10/3/035001.

Leduc, M., A. Mailhot, A. Frigon, J. Martel, R. Ludwig, G.B. Brietzke, M. Giguère, F. Brissette, R. Turcotte, M. Braun, and J. Scinocca, 2019: The ClimEx Project: A 50-Member Ensemble of Climate Change Projections at 12-km Resolution over Europe and Northeastern North America with the Canadian Regional Climate Model (CRCM5). J. Appl. Meteor. Climatol., 58, 663–693, https://doi.org/10.1175/JAMC-D-18-0021.1

Lenderink, G., B.J.J.M. van den Hurk, A.M.G. Klein Tank, G.J. van Oldenborgh, E. van Meijgaard, H. de Vries and J.J. Beersma (2014): Preparing local climate change scenarios for the Netherlands using resampling of climate model output. Environmental Research Letters, **9**, 11, 115008, doi:10.1088/1748-9326/9/11/115008

Lucas-Picher P, Caya D, de Elia R, Laprise R (2008) Investigation of regional climate models' internal variability with a ten-member ensemble of 10-year simulations over a large domain. Clim Dyn 31:927–940. doi:10.1007/s00382-008-0384-8

Madden, R.A. (1976): Estimates of the natural variability of time-averaged sea-level pressure. Mon Weather Rev 104:942– 952

Maher, N., S.B. Power, and J. Marotzke (2021): More accurate quantification of model-to-model agreement in externally forced climatic responses over the coming century. Nature Communications 12:788, doi:10.1038/s41467-020-20635-w

McSweeney, C. F., Jones, R. G., Lee, R. W., Rowell, D. P. (2015) Selecting CMIP5 GCMs for downscaling over multiple regions. Clim. Dyn. DOI10.1007/s00382-014-2418-8.

Milinski, S., Maher, N., and Olonscheck, D.: How large does a large ensemble need to be?, Earth Syst. Dynam., 11, 885–901, https://doi.org/10.5194/esd-11-885-2020, 2020

Philip, S.Y., S.F. Kew, G.J. van Oldenborgh, E. Aalbers, R. Vautard, F.E.L. Otto, K. Haustein, F. Habets and R. Singh (2018): Validation of a Rapid Attribution of the May/June 2016 Flood-Inducing Precipitation in France to Climate Change J. Hydrometeor., 19, 1881-1898, doi:10.1175/JHM-D-18-0074.1

Sieck, K. and Jacob, D. (2016) Influence of the Boundary Forcing on the Internal Variability of a Regional Climate Model. *American Journal of Climate Change*, **5**, 373-382. doi: 10.4236/ajcc.2016.53028.

Vautard, R., van Oldenborgh, G. J., Otto, F. E. L., Yiou, P., de Vries, H., van Meijgaard, E., Stepek, A., Soubeyroux, J.-M., Philip, S., Kew, S. F., Costella, C., Singh, R., and Tebaldi, C. (2019): Human influence on European winter wind storms such as those of January 2018, Earth Syst. Dynam., 10, 271–286, https://doi.org/10.5194/esd-10-271-2019

Vogel, M. M., J. Zscheischler, and S.I. Seneviratne (2018): Varying soil moisture–atmosphere feedbacks explain divergent temperature extremes and precipitation projections in central Europe, Earth Syst. Dynam., **9**, 1107–1125, https://doi.org/10.5194/esd-9-1107-2018

von Trentini F, Leduc M, Ludwig R (2019): Assessing natural variability in RCM signals: comparison of a multi model EURO-CORDEX ensemble with a 50-member single model large ensemble. Climate Dynamics. doi: 10.1007/s00382-019-04755-8

von Trentini, F., Aalbers, E. E., Fischer, E. M., and Ludwig, R. (2020): Comparing interannual variability in three regional single-model initial-condition large ensembles (SMILEs) over Europe, Earth Syst. Dynam., **11**, 1013–1031, https://doi.org/10.5194/esd-11-1013-2020

8. Appendix

The figures A1-A4 repeat the figures 2-5, but for the MPI_ESM-LR (M2) driven ensembles.

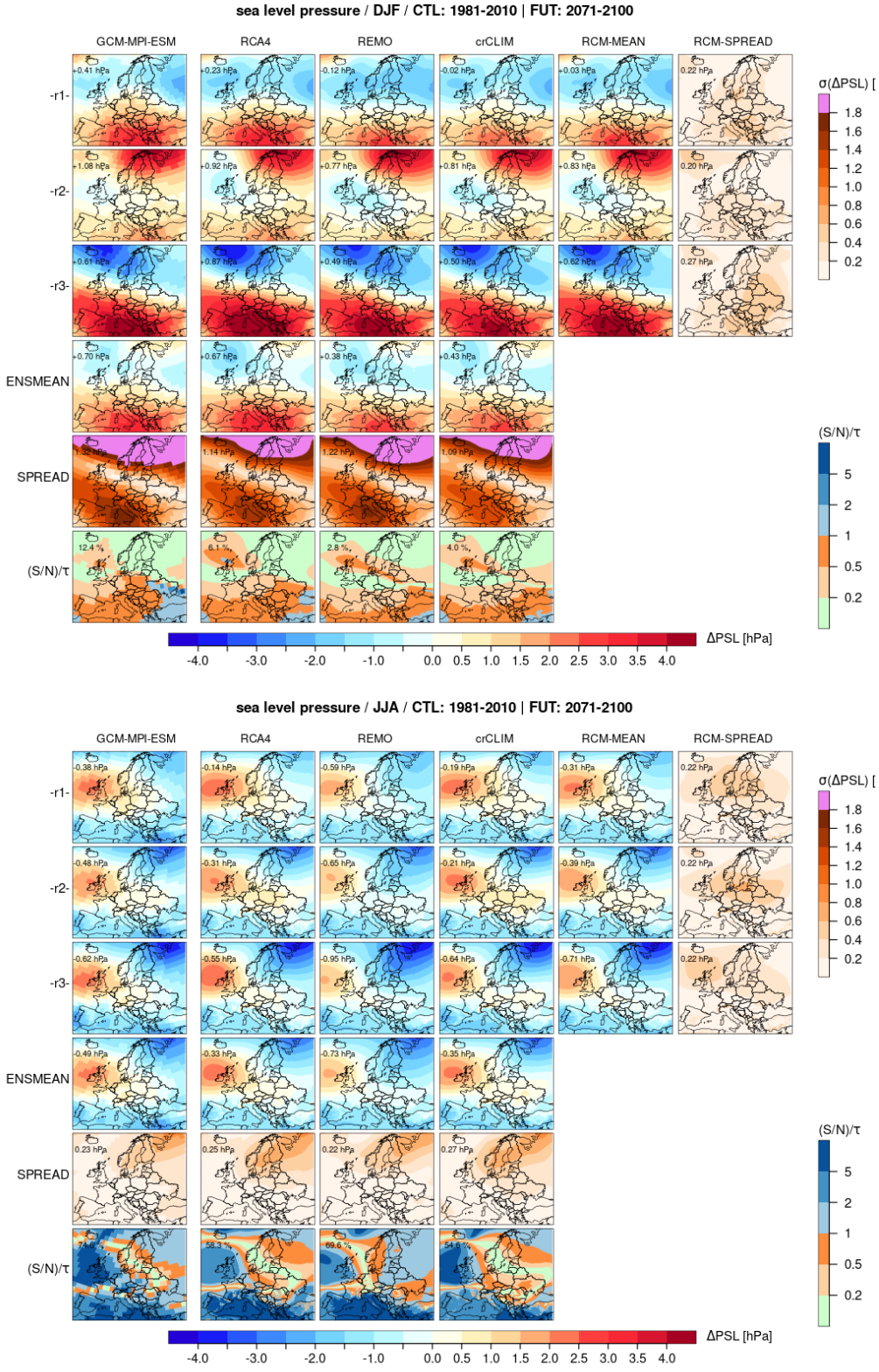

Fig A-1 Same as Fig 2 but for M2-driven ensemble. DJF (top panel); JJA (bottom panel)

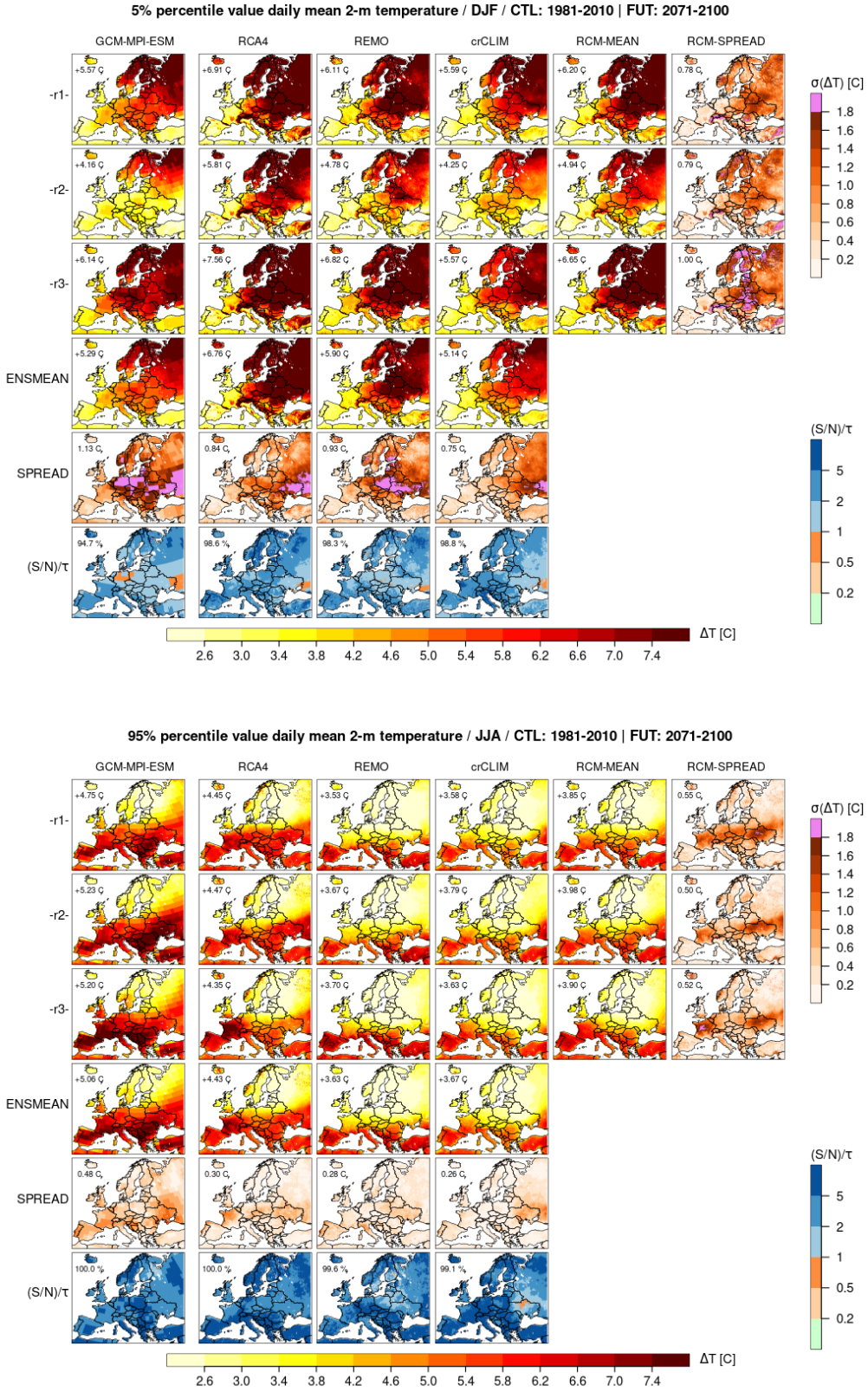

Fig A-2 Same as Fig 3 but for M2-driven ensemble. DJF (top panel); JJA (bottom panel)

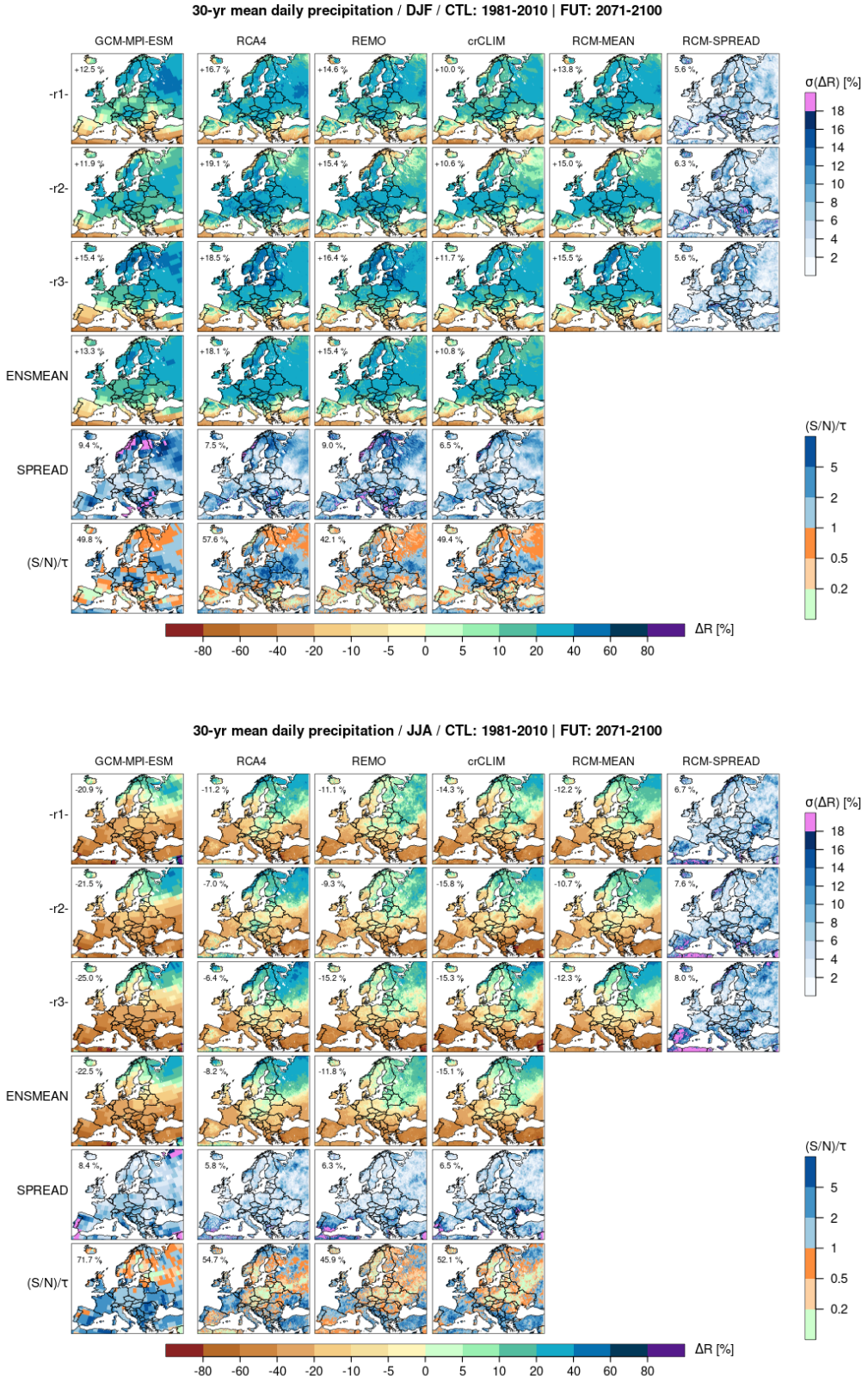

Fig A-3 Same as Fig 4 but for M2-driven ensemble. DJF (top panel); JJA (bottom panel)

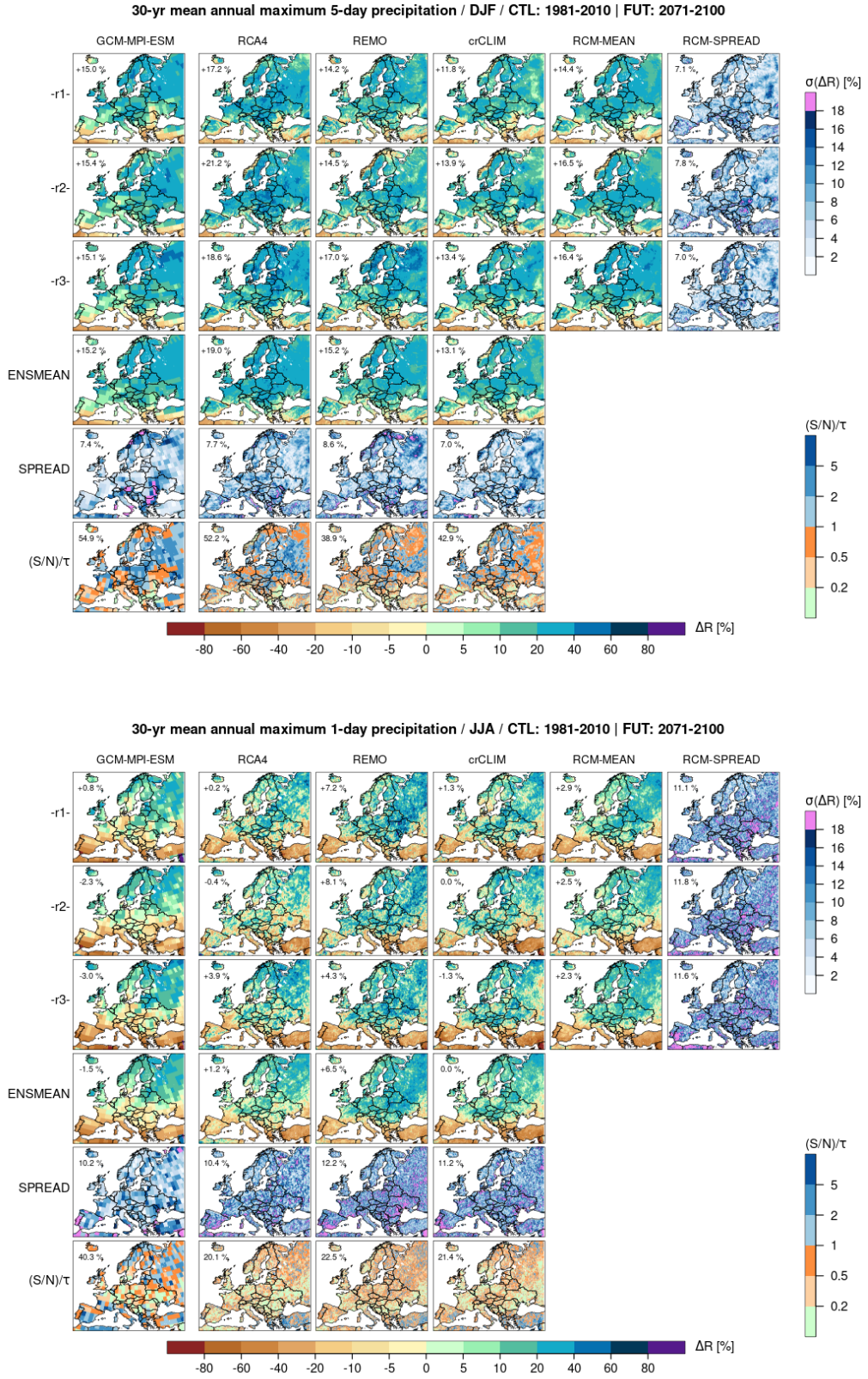

Fig A-4 Same as Fig 5 but for M2-driven ensemble. DJF (top panel); JJA (bottom panel)