Earth Syst. Dynam., 11, 1013–1031, 2020
https://doi.org/10.5194/esd-11-1013-2020

Earth System
Dynamics
Open Access

Comparing interannual variability in three regional single-model initial-condition large ensembles (SMILEs) over Europe

Fabian von Trentini[1], **Emma E. Aalbers**[2,3], **Erich M. Fischer**[4], **and Ralf Ludwig**[1]

[1]Department of Geography, Ludwig-Maximilians-Universität, Munich, 80333, Germany
[2]Royal Netherlands Meteorological Institute (KNMI), P.O. Box 201, 3730 AE De Bilt, the Netherlands
[3]Institute for Environmental Studies (IVM), Vrije Universiteit, Amsterdam, 1081 HV, the Netherlands
[4]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, 8092, Switzerland

Correspondence: Fabian von Trentini (fabian.trentini@lmu.de)

Received: 8 November 2019 – Discussion started: 20 November 2019
Revised: 24 August 2020 – Accepted: 17 September 2020 – Published: 18 November 2020

Abstract. For sectors like agriculture, hydrology and ecology, increasing interannual variability (IAV) can have larger impacts than changes in the mean state, whereas decreasing IAV in winter implies that the coldest seasons warm more than the mean. IAV is difficult to reliably quantify in single realizations of climate (observations and single-model realizations) as they are too short, and represent a combination of external forcing and IAV. Single-model initial-condition large ensembles (SMILEs) are powerful tools to overcome this problem, as they provide many realizations of past and future climate and thus a larger sample size to robustly evaluate and quantify changes in IAV. We use three SMILE-based regional climate models (CanESM-CRCM, ECEARTH-RACMO and CESM-CCLM) to investigate downscaled changes in IAV of summer and winter temperature and precipitation, the number of heat waves, and the maximum length of dry periods over Europe. An evaluation against the observational data set E-OBS reveals that all models reproduce observational IAV reasonably well, although both under- and overestimation of observational IAV occur in all models in a few cases. We further demonstrate that SMILEs are essential to robustly quantify changes in IAV since some individual realizations show significant IAV changes, whereas others do not. Thus, a large sample size, i.e., information from all members of SMILEs, is needed to robustly quantify the significance of IAV changes. Projected IAV changes in temperature over Europe are in line with existing literature: increasing variability in summer and stable to decreasing variability in winter. Here, we further show that summer and winter precipitation, as well as the two summer extreme indicators mostly also show these seasonal changes.

1 Introduction

In addition to the changes in mean climatological states, the variability of the climate system is an important feature of climate change. This variability of the climate system is subject to various drivers. Variability can be caused by natural forcings on different timescales, such as changes in solar radiation or volcanic eruptions. Variability of single components of the climate system can also be caused by the redistribution of heat and momentum between and within different components (e.g., ocean and atmosphere) of the coupled climate system, referred to as unforced internal variability. Next to these variations, anthropogenic changes in greenhouse gas concentrations contribute to a changing climate. Climate variability can be sampled on different timescales from hours and days up to multi-decadal variations.

For impact analysis of climate change, the future development of interannual variability (IAV) is of utmost importance in addition to changes in the mean climate state. Particularly increases in the IAV can be crucial for many impact sectors, as this makes it much harder for stakeholders to plan from year to year. In this study, daily data are used to calculate six indicators (summer and winter mean surface air temperature (tas) and accumulated seasonal precipitation (pr)) and two indicators for climatological extremes with high societal impact: the number of heat waves per year (tas-HW-Nr) and the maximum length of dry periods per year (pr-DP-MAX); see Table 1 for definitions. Heat waves can cause an increase in health problems and even fatalities among the population, as well as damage to infrastructure (e.g., highways) and ecological problems, as seen during the most recent heat waves in Europe (e.g., 2003, 2018, 2019). Long dry periods can have major impacts on ecology, forestry, agriculture, drinking water supply, power plant cooling outages, transport on rivers and many more. All these sectors should implement adaptation strategies to face changing climatic conditions, including IAV.

Early studies with regional climate models from PRUDENCE showed a distinct increase in IAV for the 21st century in summer temperatures (Fischer and Schär, 2009, 2010; Vidale et al., 2007) as well as decreasing winter temperature variability (Vidale et al., 2007). Later work with ENSEMBLES models revealed a less pronounced increase in summer temperature variability (Fischer et al., 2012). Analysis of SMILEs also showed future increases in variability of European summer temperatures with increasing global warming (Suarez-Gutierrez et al., 2018; Yettella et al., 2018). Holmes et al. (2016) and Tamarin-Brodsky et al. (2020) also find increasing temperature variability in summer and decreasing variability in winter for the future. European winter temperature variability has already today decreased since the pre-industrial era in another large climate model ensemble (Bengtsson and Hodges, 2019).

For large areas of the globe, including Europe, an increase in precipitation variability from daily to multi-decadal timescales is expected due to higher temperatures (Pendergrass et al., 2017). However, Ferguson et al. (2018) find significant changes in the IAV of monthly precipitation only in a small fraction of CMIP5 models for a western European domain until the end of the 21st century. Earlier analysis with regional climate models revealed future increases in summer and decreases in winter for IAV of precipitation over similar domains as used in this study (Giorgi et al., 2004).

Uncertainty of future climate projections can stem from at least three sources (Hawkins and Sutton, 2009): emission scenario, model response to a selected forcing and internal variability of the climate system. Internal variability is often referred to as "irreducible uncertainty" at timescales beyond seasons to decades. While scenario and model response uncertainty have been referred to in many climate simulation experiments (CMIP and CORDEX), the internal variability component had received less attention for many years. In recent years, a new tool for the assessment of internal variability has become quite popular: single-model initial-condition large ensembles (SMILEs), where the same model is forced with the same emission scenario several times – with the runs (members) just differing in their initial conditions. This setup is able to isolate the internal variability component from the scenario and model response uncertainty for the respective model. Based on SMILEs, it has been shown that the contribution of internal variability to the total uncertainty of multi-model ensembles (CMIP, CORDEX) can be large, especially for mid-term projections and precipitation (Kumar and Ganguly, 2018; von Trentini et al., 2019) at the regional level.

The terms large ensemble (LE) and SMILE usually describe the same thing, but we prefer SMILE as it incorporates the type of large ensemble which is built upon different initial conditions. Up to now, a number of large ensembles have been produced. Deser et al. (2020) give the latest overview of the different SMILEs available. However, most studies only use one SMILE for their analysis and the rare comparisons are usually just between two ensembles: similar patterns of internal variability of temperature and precipitation trends for the middle of the 21st century were found for a CCSM3 and an ECHAM5 ensemble over North America by Deser et al. (2014). Martel et al. (2018) showed a consensus of the IAV of annual mean and extreme precipitation in a CanESM2 large ensemble (which is also used for boundary conditions of the CRCM5 in this study; see Sect. 2) and CESM-LE with two global observational data sets.

All these simulations were performed with global climate models (GCMs), and only a few were dynamically downscaled with regional climate models (RCMs). Here, we compare three dynamically downscaled large ensembles, all forced by the Representative Concentration Pathway (RCP) 8.5, for Europe. It is the first time that regional large ensembles are compared with respect to forced changes and their internal variability. The added value of RCM simulations is well documented for EURO-CORDEX (Giorgi et al., 2009; Torma et al., 2015; Sørland et al., 2018; Giorgi, 2019). Downscaled climate data are also a necessity for impact modeling at regional to local scales (e.g., for hydrology, agriculture, biodiversity research) due to their more accurate representation of topography, complex coastlines and the heterogeneity of land surface properties.

Terminology in the context of climate variability is not always clear in the literature, as the terms natural variability, internal variability, IAV and inter-member variability (IMV) are often used synonymously or mixed up. Here, the term internal variability describes the variability at timescales from seconds up to multiple decades caused by unforced internal effects of a model or the real world due to the chaotic nature of the climate system only, without incorporating naturally forced variability due to volcanic eruptions and solar forcing. Anthropogenic changes in greenhouse gas and aerosol concentrations as well as natural solar and volcanic forcing can cause changes in the mean climate state that are superimposed by internal variability. Moreover, higher greenhouse gas concentrations can cause changes in the internal variability itself in the future as well – adding another component to climate change effects.

Table 1. Indicators and their definitions.

Indicator	Used variable	Definition
tas-JJA	tas	Summer mean temperature (June–August)
tas-DJF	tas	Winter mean temperature (December–February)
tas-HW-Nr	tas	Number of heat waves per year; a heat wave is defined as a minimum of 3 d above the 95th percentile of daily mean temperature of the reference period; no filtering on summer months is applied at any stage; however, by definition the heat waves will occur during summer in the reference period. They might, however, extend to spring and fall under RCP8.5.
pr-JJA	pr	Summer precipitation sum (June–August)
pr-DJF	pr	Winter precipitation sum (December–February)
pr-DP-MAX	pr	Maximum length of a dry period per year; dry periods are a minimum of 11 consecutive days with every day showing less than 1 mm of precipitation; no filtering on summer months is applied at any stage; the periods can thus also occur in winter (but this is rather unlikely in Europe).

In this study, IAV is calculated as the standard deviation of anomalies of each member from the ensemble mean (EM), which represents an estimate of the forced response of the respective model. SMILEs have the advantage that the EM is a much better estimate of the forced response than a trend fitted to single members. After removing the forced response, the residual IAV equals the total unforced internal variability – including low-frequency variations. We will show that IAV can be well estimated by the IMV of a SMILE in many cases, as both metrics sample the unforced internal variability of a SMILE, just on different dimensions: IAV is sampled on the time dimension of a single member, while IMV is sampled on the member dimension for each year (see Sect. 3.4). Both IAV and IMV are terms used to describe the more general term "internal variability" throughout this paper. Also, note that the EM of each SMILE can be regarded as the change signal with the highest probability, but which specific member would become realized depends on internal variability.

The usage of three RCM-SMILEs has some advantages compared to multi-model ensembles consisting of single realizations that enable us to go beyond the current literature on IAV changes. First, we can better evaluate the models against observations as (a) the forced response of the model is better estimated by the EM and (b) we thus reduce the problem of having only one realization of climate to the observational data side of the evaluation. Second, we can more reliably quantify changes in the IAV and rule out that potential changes only occur as a result of sampling uncertainty. Additionally, we can better demonstrate when changes are significantly different from historical conditions. In recent literature, often no significance test of detected changes in interannual variability is performed (e.g., Bengtsson and Hodges, 2019). Many studies just provide information about the robustness of change (e.g., by stippling in maps), measured by the accordance in the sign of change of (usually) 67 % of the models of multi-model ensembles (e.g., Holmes et al., 2016). This does, however, not allow information about the significance compared to a reference climate. Third, SMILEs allow a better separation of models as they are not only described by one member each. Additionally, we combine these general SMILE advantages with the higher resolution of RCMs.

The remaining paper is structured as follows: First, the model ensembles and the observational data set E-OBS are briefly presented. Then, the change in mean temperature and precipitation together with the inter-member spread of projected changes is analyzed for each ensemble, as the mean changes are important baseline information for variability changes. Next, the IAV of the three regional large ensembles is evaluated against E-OBS to assess the abilities of the models to represent observed IAV for the selected indicators. Finally, IAV in historical climate and future changes in IAV are compared between SMILEs. This includes a discussion on different methods to estimate IAV and detect significant changes in IAV. In the main text, most results will only be presented for mid-Europe (ME), with references to the other regions and their figures in the Supplement.

2 Data

The climate model ensembles each consist of a GCM single-model initial-condition large ensemble, which has been dynamically downscaled over Europe with a single regional climate model: a 50-member CanESM2-CRCM5 ensemble (Kirchmeier-Young et al., 2017; Leduc et al., 2019), a 21-member CESM-CCLM ensemble (Fischer et al., 2013; Addor and Fischer, 2015; Brönnimann et al., 2018) and a

Table 2. Specifications of the three ensembles used in this study.

	CRCM	CCLM	RACMO
Scenario	RCP8.5	RCP8.5	RCP8.5
GCM	CanESM2	CESM 1.0.4	EC-EARTH 2.3
GCM resolution	2.8°	2.0°	1.0°
RCM	CRCM5	CCLM4-18-7	RACMO22E
RCM resolution	0.11°	0.44°	0.11°
No. of members	50	21	16

16-member EC-EARTH-RACMO ensemble (Aalbers et al., 2018), all forced with the RCP8.5 scenario, resolved on different spatial resolutions (Table 2). Hereafter we indicate the GCM-RCM combinations with the RCM names only (CRCM, RACMO and CCLM). This setup with a shared scenario but different models enables us to analyze differences in internal variability between the three ensembles. Differences in variability may stem from the differences in the resolution of both GCMs and RCMs, the different domain sizes, the different models, and differences in aerosol forcing in the RCM simulations (constant in CCLM and CRCM, transient in RACMO) and in the application of an ocean slab model in the EC-EARTH-RACMO ensemble. RACMO also uses slightly different grid specifications. The domain size of CCLM equals the EURO-CORDEX domain, while CRCM uses a slightly smaller domain, and RACMO only captures central and northeastern Europe (Fig. 1). The initialization is carried out differently in the three driving GCM-SMILEs: CanESM2 builds on a hybrid approach, where five members with different ocean conditions starting in 1850 were divided into 10 members each using atmospheric perturbations during the initialization in 1950 (see Leduc et al., 2019 for details). The CESM members stem from small atmospheric perturbations of the order of 10^{-13} on 1 January 1950 (Fischer et al., 2013). EC-EARTH uses the first 16 d in the year 1850 of an initial run to start the 16 members (Aalbers et al., 2018). These climate model data sets will be compared but will also be compared to observations: the gridded observational data set E-OBS has daily precipitation and temperature available for Europe (version v12.0, spatial resolution of 0.22° on a rotated pole grid). We use the E-OBS data set because of its availability for Europe and it has similar spatial resolution to the regional climate models under consideration. We accept the known weaknesses of the data set (mostly caused by inhomogeneities in the sparse station network; E-OBS is also known for rather low precipitation fields; see Hofstra et al., 2009) and assume that it is nonetheless suitable for the purpose of this study.

3 Methods and results

3.1 Spatial aggregation

All indicators (Table 1) are calculated on a grid basis for each ensemble. For comparison, the indicators are spatially aggregated to four regions in Europe, for which all three RCM domains overlap (Fig. 1): the British Isles (BI), France (FR), mid-Europe (ME) and the Alps (AL). These regions are well known from other European climate model studies (Lenderink, 2010; Lorenz and Jacob, 2010; Kotlarski et al., 2014; von Trentini et al., 2019) and were introduced by Christensen and Christensen (2007). The procedure of calculating the indicators on the grid level and spatially aggregating them afterwards has the advantage that no regridding of data is needed. However, the different spatial resolutions of the models alone can potentially lead to higher variability in the 0.11° data (CRCM and RACMO), compared to the 0.22° (E-OBS) and 0.44° (CCLM) data. This is especially the case for spatially heterogeneous variables and indicators (Giorgi, 2002; Kendon et al., 2008). The indicators in this study, however, have relatively low spatial heterogeneity (seasonal temperature and precipitation, heat waves, and dry periods are rather large-scale phenomena), where the range of spatial resolutions of the data used here (between 0.11 and 0.44°) is not expected to be significantly sensitive. The effect of regridding before the calculation of indicators is shown by a short experimental analysis, where 1 year of five members of the 0.11° CRCM data is regridded to 0.44° (simply averaging 4 × 4 grid cells each), before the indicators are calculated. The results show that the effect of regridding on the IMV is indeed minor for the indicators considered (Fig. S1). The approach of direct regional aggregation of the indicators calculated on the grid level is therefore applied for the further analysis of this study.

3.2 Ensemble spread of projected mean climate change

Before analysis of IAV, simple scatterplots of the changes in the mean climatological states of each member for temperature and precipitation for summer and winter between 1980–2009 and 2070–2099 are shown for ME; see Figs. 2 and 3. They give a first impression on the spread of projected changes between the members of SMILEs and on the differences in the mean changes between models. We test the similarity of means between all models with a two-sample t test ($\alpha = 0.05$) and the similarity of spreads with a Brown–Forsythe test (BF test with $\alpha = 0.05$) on equal variances. The BF test does not show significant differences in variance for both temperature and precipitation in winter and summer for all model combinations. The spread of signals between members of one SMILE can be solely attributed to the internal variability of the respective model.

In summer in ME, all models show decreasing precipitation; between −3 % and −16 % for RACMO and −14 % to

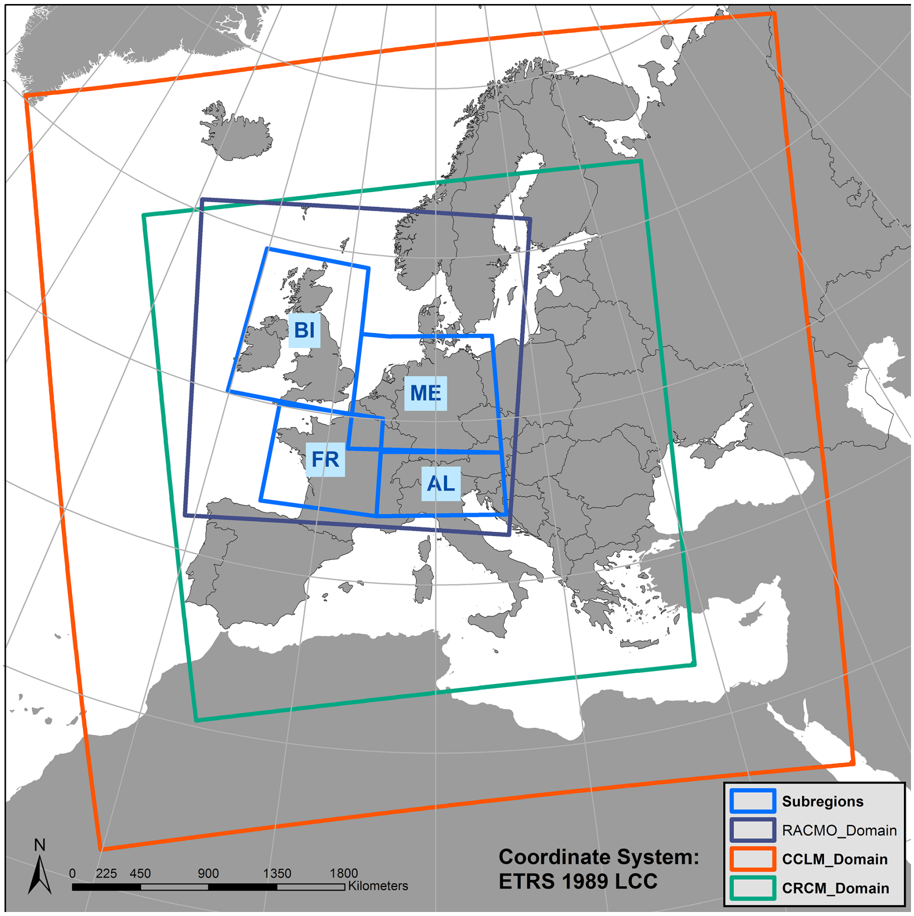
BI
ME
FR
AL
N
0 225 450 900 1350 1800 Kilometers
Coordinate System: ETRS 1989 LCC
Subregions
RACMO_Domain
CCLM_Domain
CRCM_Domain

Figure 1. Domains of the three RCMs and the boundaries of the four analysis regions; BI: the British Isles; FR: France; ME: mid-Europe; AL: Alps; the CCLM domain matches the EURO-CORDEX domain.

−35 % for CRCM and CCLM (Fig. 2). Increases in summer temperature between 3 and 5 °C are projected by RACMO and CCLM, while CRCM shows much higher changes between 5 °C and more than 6 °C. Thus, RACMO and CCLM show similar changes in temperature (although statistically different in their means), while CCLM and CRCM show similar changes in precipitation. The spread of changes for both temperature and precipitation of RACMO and CRCM are similar, both in terms of standard deviation and total range, while CCLM shows a higher standard deviation and total range (Table 3). Similar results as discussed here for mid-Europe (mean changes and spread) are found for France and the Alps (not shown), with the largest decrease in summer precipitation over France and the strongest warming over the Alps. The British Isles region shows less pronounced changes for both temperature and precipitation (although they are consistent in sign). CRCM shows a closer similarity of precipitation decreases to RACMO in BI rather than to CCLM, as is the case for the other three regions ME, FR and AL.

In winter, all models project increasing precipitation (1 %–32 %) and temperature increases between 1.4 and 5 °C by the end of the 21st century (Fig. 3). RACMO and CRCM show a similar standard deviation and range of temperature and precipitation changes again, together with similar mean changes as well (significant for temperature but not for precipitation). CCLM shows distinctly smaller changes in combination with a smaller spread of changes (Table 3). Similar results also appear for FR, AL and BI, although some members of CCLM and CRCM also project a slight decrease in precipitation in these regions.

3.3 Evaluation against E-OBS

For the evaluation of the models' IAV against E-OBS, we apply an approach proposed by Suarez-Gutierrez et al. (2018) and Maher et al. (2019). For the observations and for each model and member separately, the anomalies relative to the reference period 1961–1990 are calculated for the years 1957–2099 and 1957–2015 in E-OBS, respectively (Fig. 4). Model mean state biases of the indicators, which can be quite large (see Fig. S2) are thereby removed. For each year,

Table 3. Standard deviation and total range for changes in Figs. 2 and 3.

	tas (°C)			pr (%)		
	CRCM	RACMO	CCLM	CRCM	RACMO	CCLM
Summer SD	0.27	0.28	0.39	4.2	3.8	5.1
Summer range	1.16	0.96	1.59	16.2	13.2	21.1
Winter SD	0.37	0.38	0.30	5.5	5.3	4.2
Winter range	1.87	1.47	1.33	26.1	21.2	12.9

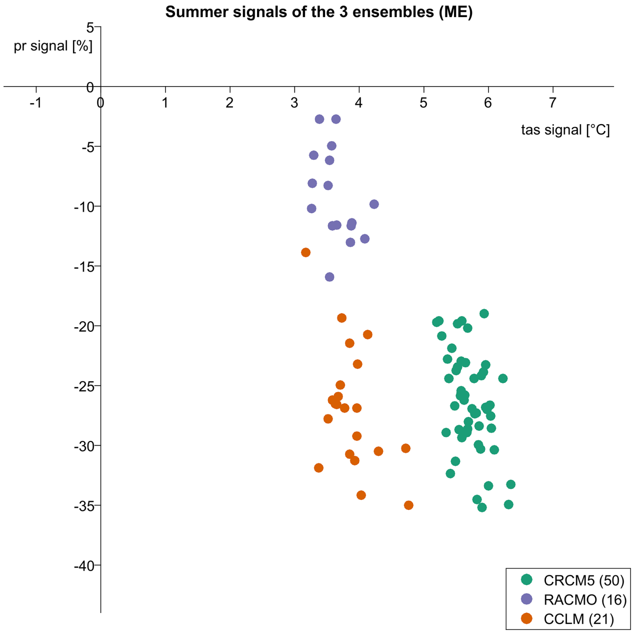

Figure 2. Change in mean summer temperature and precipitation for every member of the three ensembles in mid-Europe (2070–2099 against 1980–2009). Changes are relative to each members' value in 1980–2009 for precipitation, while temperature changes are absolute.

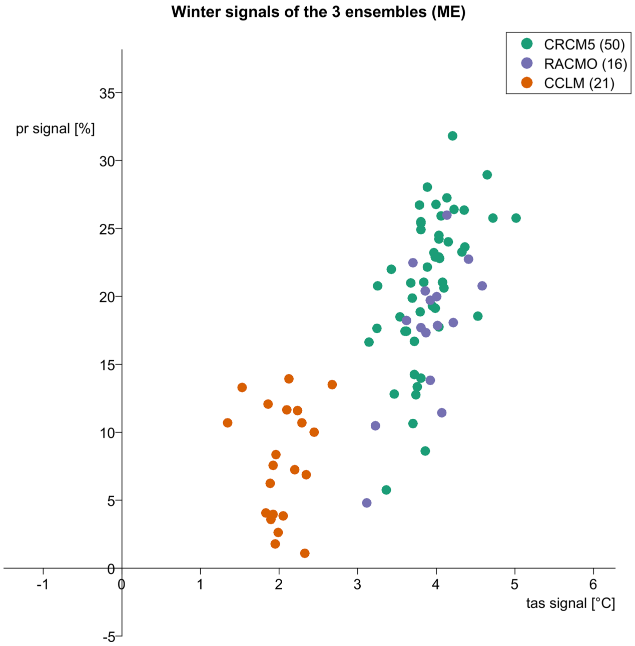

Figure 3. Change in winter temperature and precipitation for every member of the three ensembles in mid-Europe (2070–2099 against 1980–2009). Changes are relative to each members' value in 1980–2009 for precipitation, while temperature changes are absolute.

we then plot the ensemble median, minimum and maximum member, the area between the 12.5th and 87.5th percentile, within which 75 % of the members are situated, and the E-OBS data. For a perfect model, the E-OBS data are expected to occur normally distributed within the range spanned by the ensemble and are concentrated in the inner 75 %, several years situated in between the minimum and maximum of members, but also outside this range from time to time. If the E-OBS data are concentrated too much inside the total range or even the 75 % area, the variability of the ensemble overestimates the observational variability. By contrast, if too many E-OBS data points exceed the ensemble spread, SMILE underestimates observational variability. To quantify this further, the probability density function of the anomalies in the period 1957–2015 is plotted for each member and E-OBS separately. The functions are estimated probability densities based on a normal kernel function, similar to an approach by Lehner et al. (2018).

The forced response (ensemble median) increases for all indicators analyzed, except for summer precipitation, which decreases in all models, and there is no clear change in pr-DP-MAX in RACMO. Note that this approach does not only compare the IAV of the models and E-OBS, but also the forced response in the historical period. Differences in the distributions can thus also arise from a false representation of the forced response in a model, compared to the trend in E-OBS. On the other hand, if the modeled and observed distributions largely coincide, both the forced response and the IAV are well represented by a model. All three models generally seem to reproduce the forced response in the historical part quite well, as the models are consistent with the trends of the E-OBS points (e.g., increase in tas-JJA). Only for summer precipitation (pr-JJA) do all models show a decrease in

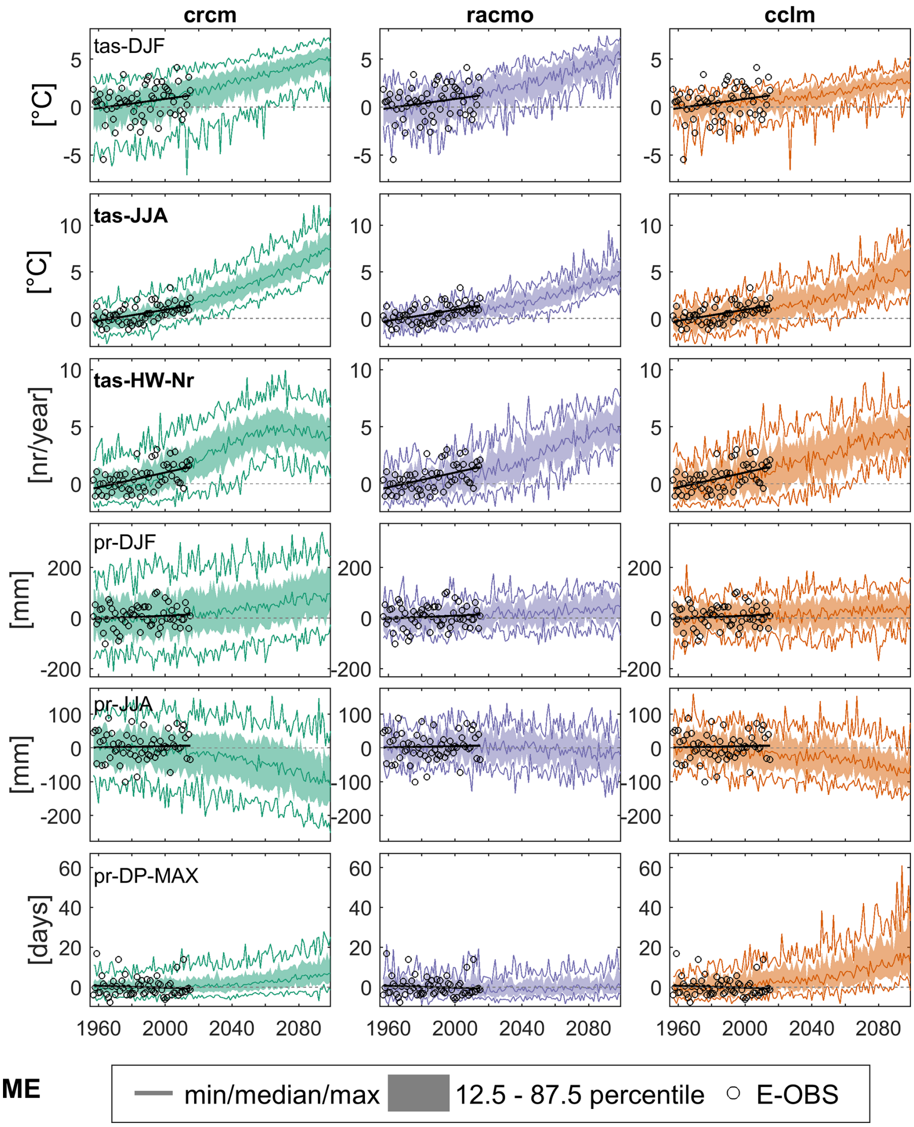

Figure 4. Anomalies from 1961 to 1990 of the six indicators in mid-Europe (ME) for E-OBS (circles 1957–2015) and the three ensembles (1957–2099), represented by the median, minimum and maximum (solid lines) of the ensemble and an area from the 12.5th and 87.5th percentile, spanning the range of the inner 75 % of the members (shading). Black lines show the linear trend for the E-OBS points. The indicator names are in bold when the trend is significant using a Mann–Kendall test ($\alpha = 0.05$).

the forced response, whereas E-OBS shows no significant negative trend. However, in all ensembles, not all members show decreasing trends. The observations may thus still be consistent with the simulated forced response.

The comparison of E-OBS and the three SMILEs during the historical period from 1957 from 2015 in ME shows largely good representations of IAV in the ensembles, as seen by well distributed E-OBS points within the 75 % range (12.5 %–87.5 % quantile) and minimum and maximum range of the ensembles (Fig. 4). However, too a strong clustering of the E-OBS points in the 75 % area occurs for winter precipitation in CRCM (97 % fall inside) and number of heat waves in CCLM (90 %), meaning the simulated IAV is too high. On the other hand, too many outliers beyond the minimum and maximum members appear in winter temperature in CCLM (22 % outside of total range), winter precipitation in RACMO (17 %) and maximum duration of dry periods in CRCM (10 %), i.e., for these models and indicators the simulated IAV is too low. To demonstrate this further we calculate probability density functions of the annual anomalies for each member and E-OBS (Fig. 5). Note that probability density functions could also be somewhat inflated by the underlying mean trend, but we expect this effect to be small because trends in the observational period are small and largely consistent between models and observations. To evaluate the ability of SMILEs to represent observational IAV, we test whether the E-OBS distribution looks like a possible member of the respective ensemble. The observations are not expected to fall near the ensemble median but rather should be ideally indistinguishable from a random additional member of the ensemble, since E-OBS only represents one possible realization of historical climate. Significant differences can be seen for the examples already mentioned: the distribution of CRCM in winter precipitation is much broader than the E-OBS distribution, whereas the winter temperature distribution for CCLM is concentrated too much in the middle compared to E-OBS.

Similar results as in ME can be found in the other three regions as well (Figs. S3–S5), with only a few cases where the E-OBS distributions show a distinctly different shape than all members of the ensembles (Fig. 6 for AL and Figs. S6 and S7 for BI and FR, respectively), especially for the maximum duration of dry periods of CCLM in France. This is not too surprising, as the maximum duration of dry periods is an extremely sensitive indicator because of its potentially extreme differences in magnitude between (model) years/members (one wet day can make a huge difference). The other two SMILEs are able to represent the E-OBS variability for this indicator in France though. Other remarkable features are the underestimation of variability in RACMO for all six indicators in the British Isles region (Fig. S6), as well as the relatively good performance of the models for the Alps (Fig. 6), which is probably the most difficult region for a model to represent correctly due to the strong spatial heterogeneity. However, the Alps show some "outlier-members" with distinctly different distributions in the ensembles (e.g., winter precipitation in CRCM), which cannot be found in the other regions – at least not this pronounced. These outliers demonstrate how large the influence of internal variability between members can be in a single realization of climate, as these outlier members just deviate from all other members by their initial conditions. Estimating the IAV of a model thus needs a large number of members, as even with 49 members that give a uniform range of distributions (pr-DJF in CRCM5 in the Alps, Fig. 6), one single additional member can change the picture and add more information on the range of IAV for the respective model. The evaluation of E-OBS gets rather difficult in these cases, as the methodology is based on the assumption that the E-OBS distribution should somehow "fit" to the uniform range of distributions of the model. If the E-OBS data showed such an outlier behavior, it means that the one realization of climate variability as seen by the E-OBS data might still be part of a SMILE's range of possible variability manifestations. However, from a probability perspective, the conclusion of similar variabilities becomes rather unlikely. It just makes it harder to prove that E-OBS has a different distribution than all members of a SMILE.

3.4 Projected changes in internal variability and the connection between IAV and IMV

The temporal development of the internal variability is important information along with the underlying forced response (change in the EM) for a better understanding of changing climatic conditions. We discuss three possible ways to describe changes in the internal variability on annual timescales within a SMILE. All three methods are based on the application of a BF test on equal variances. While IAV and IMV are expressed as standard deviations (SDs), the BF test analyses differences in the variance, which is just the square of SD. In the cases of IAV (methods 1 and 3), moving time periods of 30-year length, shifted by 1 year each (1961–1990, 1962–1991, . . . , 2070–2099) are used. For the second method, IMV is sampled over the dimension of the ensemble size per year. Thereby we test whether the internal variability changes significantly over time. Differences in the methods arise from the different data samples used for the testing.

The first method is based on the methodology that one would choose for single members and observations. By looking at the IAV for different periods within one member, changes in IAV can be detected. Usually the forced response is taken out of the data by fitting a polynomial to the data and only using the residuals. However, the estimate of the forced response of a model based on only one member may deviate from the true forced response (Lehner et al., 2020). Therefore, we choose the EM as an estimate of the forced response and use the residuals from each member with respect to the EM for the BF test. The BF test results is Boolean information for each member and each moving period on whether the variance has significantly changed with respect to a reference period (here: 1961–1990) or not. This information can be used to show the percentage of members with a significant change (separated for positive and negative changes) in each period. The advantage of SMILEs within this approach is the better estimate of the forced response and a more robust detection of changes, as they are built on multiple members. One member alone could be an outlier in its representation of (changes in) IAV just as it could be for the trend. The method is sensitive to the chosen reference period of course, as the variance of this period determines the baseline variance. Since we use moving periods, the results do not change significantly when using different periods starting in the 1960s.

For the second method, we make use of the assumption that the IMV for a given year is a good approximation for the IAV in a period around that year (e.g., ±15 years to get a sample size of the typical 30 years for climate analysis). The sampling of variability is thus not based on consecutive time series within each member but on a compound of annual data for 1 year from all members of a SMILE. The IMV is also based on residuals from the EM as for IAV. Under the assumption of small influence of low-frequency variability, IMV should be a good estimate of total unforced IAV, as

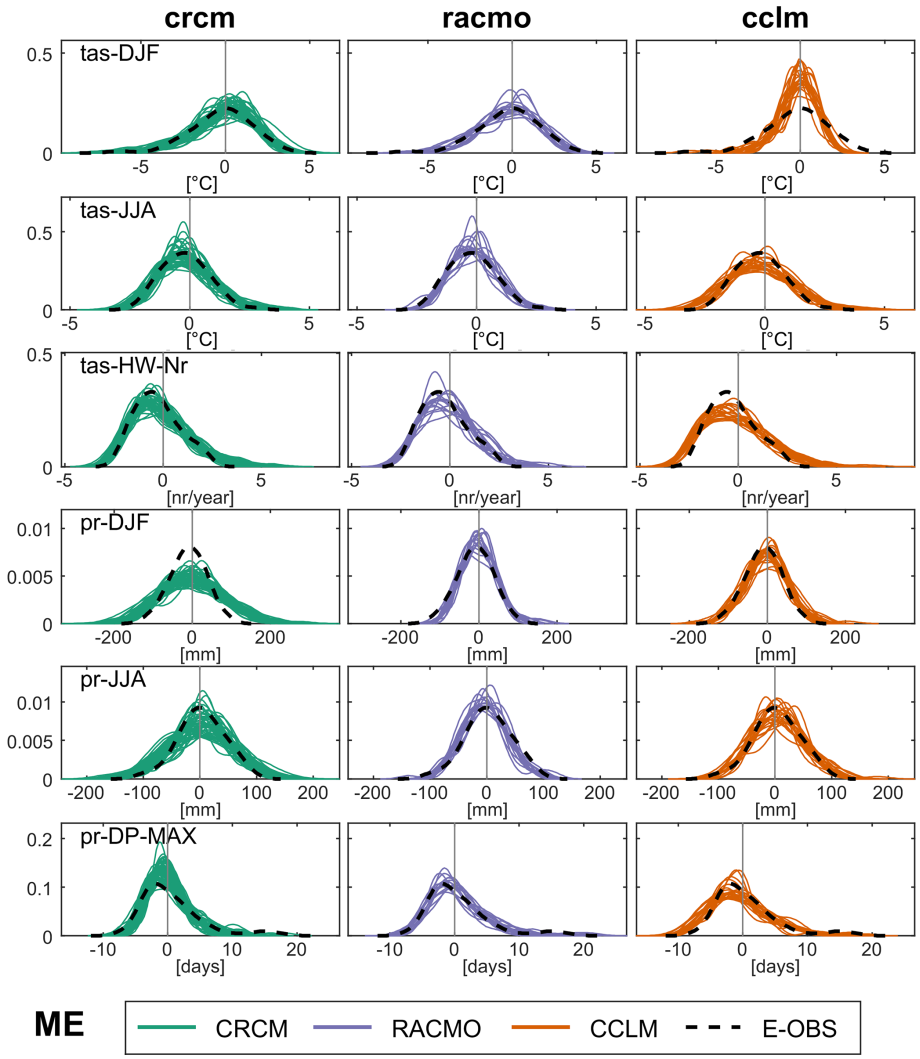

Figure 5. Probability density functions of the annual anomalies during the period 1957–2015 in E-OBS and each ensemble member for all six indicators in mid-Europe (ME).

both sample the annual variability during a similar state of climate for a given time horizon. This concept is particularly relevant in the presence of non-linear forcing. For instance, the response to a volcanic eruption cannot be separated easily from unforced IAV. In addition, the anthropogenic forcing since 1950 has not been linear in time. Using IMV is an elegant way to get around this challenge. Some recent publications support the concept of using IMV as an approximation of IAV, although the two have different background meanings: while IAV has a physical meaning and represents the variability of a consecutive sequence of weather phenomena, IMV is a measure of variability without a direct physical meaning (Nikiéma et al., 2018).

In Leduc et al. (2019) the authors state that "In the case of a climate system under transient forcing, the use of [IMV equation] to assess temporal variability using the inter-member spread involves weaker assumptions than calculating the residual temporal variability from detrended time series." (Leduc et al., 2019, p. 681), based on the study by Nikiéma et al. (2018). A recent publication by Wang et al. (2019) even concludes that the IMV of winter sea level pressure over Eurasia in a SMILE is driven by the same mechanisms as observed IAV via an EOF analysis. Another example is the analysis of seasonal mean and heavy precipitation in Europe where long-term variations are small compared to the IAV in the RACMO ensemble (Aalbers et al., 2018). A comparison of IAV and IMV in each ensemble is carried out by comparing the means and standard deviations of these two variability metrics – calculated over different dimensions of the ensemble data. The IAV is calculated for each member during a 30-year reference period (1980–2009) and three future periods. The mean and standard deviation of these 50, 21 and 16 values is calculated for IAV. The IMV is calculated for each of the 30 years of the respective period between the 50,

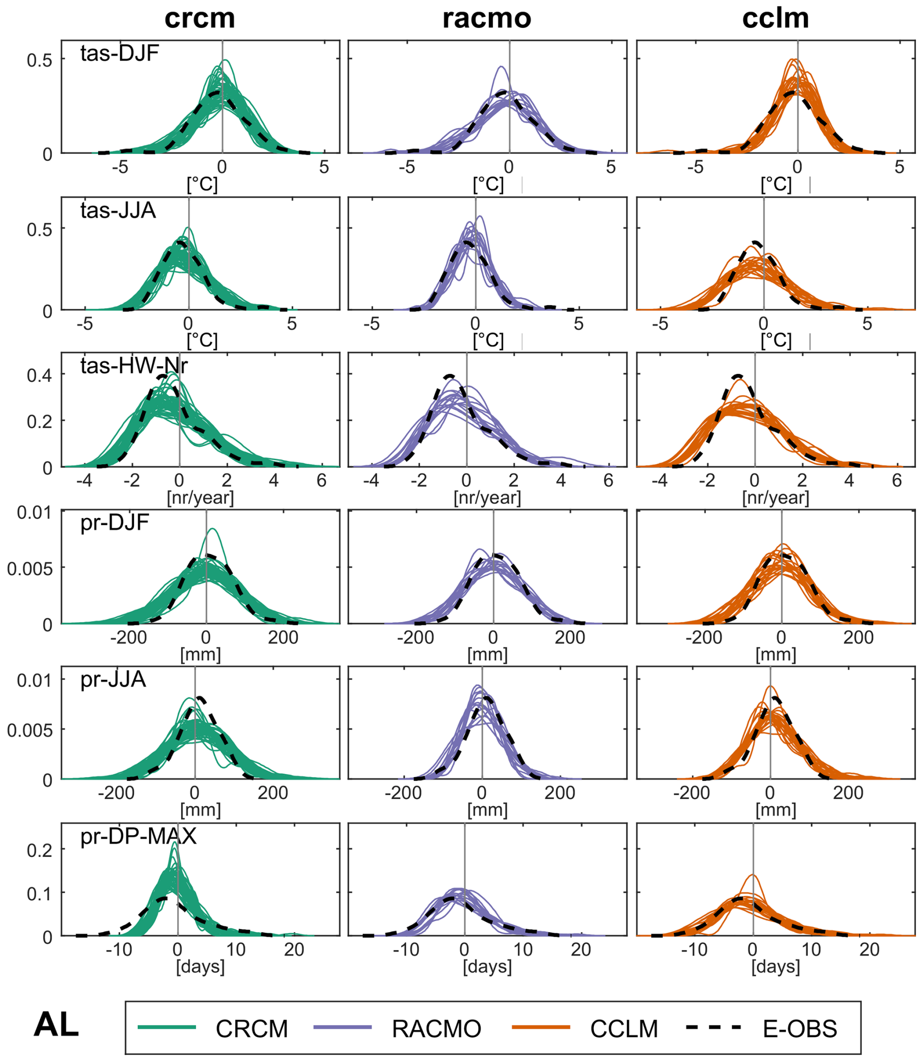

Figure 6. Probability density functions of the annual anomalies for all six indicators in the Alps (AL). For details, see Fig. 5.

21 and 16 members, leading to a mean and standard deviation, calculated from these 30 values. The mean and standard deviation of IAV and IMV are indeed very similar for all indicators, periods and regions (exemplarily shown for winter temperature in Fig. 7). Especially the similarity in future changes suggests a similar response to external forcing for the two variability metrics IAV and IMV. The IMV has the advantage that it is insensitive to inflation effects of the variability due to an existing trend and forced effects like cooling after volcanic eruptions for example. However, although IAV and IMV seem to be similar in many cases (see also literature above), they can potentially also differ under special circumstances in the external forcing like volcanic eruptions. Note that according to our results IMV is always slightly larger than IAV. This may be caused by two factors. First, detrending the time series is more likely to remove than to add some of the variability, and it affects only IAV. Second, also without detrending the data, in the presence of low-frequency variability, IAV is likely smaller than IMV, which has no auto-correlation in the underlying data. For the variables considered here, differences are small though, implying that the low-frequency variability is indeed small compared to the high-frequency variability.

Given the similarity of IAV and IMV, the third approach pools together the annual anomalies from the EM from all members for a given 30-year period (30 times the ensemble size, e.g., 30 × 21 values for CCLM). It is therefore a mixture of IAV and IMV, enabling a more robust BF test result for changes in variance by a larger sample size.

While the interpretation of temperature-based indicators is always based on absolute anomalies from the EM, it can be useful to look at both absolute and relative anomalies from the EM for precipitation-based indicators (in contrast to the previous evaluation against E-OBS, where they were only absolute anomalies). Relative anomalies thus give information on how much the standard deviation changes with respect

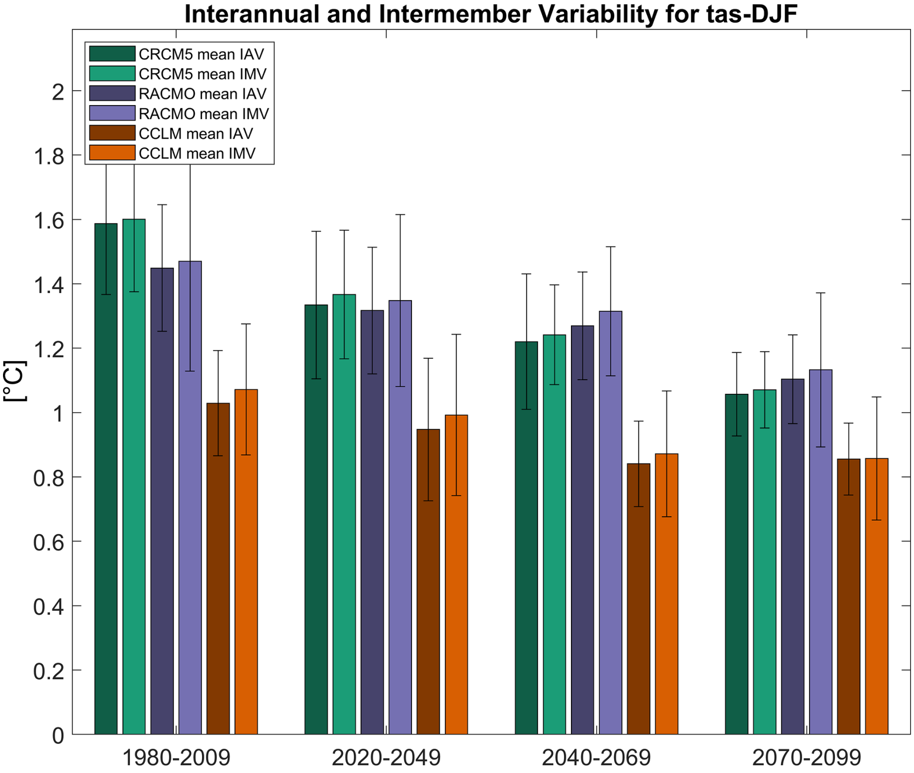

Figure 7. IAV and IMV of winter temperature in the three ensembles for the reference period (1980–2009) and three future periods. Bars: mean over the variability of each member (IAV) or year (IMV). Error bars: ± standard deviation (members or years); IMV: 16/21/50 members; IAV: 30 years of the respective period.

to changes in the EM. For example, a stable IMV in absolute terms will result in a decrease in the relative IMV when the EM increases. Increasing relative IMV, together with an increasing EM on the other hand means that the internal variability is increasing even more than the mean.

The percentage of members with significant changes in IAV as a function of time is shown in Fig. 8 for all indicators, for mid-Europe. Significantly decreasing IAV for winter temperature and increases for summer temperature and heat waves are found, but only for a minority (<50 %) of the members for all models, even at the end of the 21st century. While all three models point to the same direction of change, percentages differ substantially. For winter and summer precipitation, an even smaller percentage of members shows significant changes in IAV, and there is no clear direction of change in any model. Only CRCM for pr-JJA shows an increasing number of members with significant positive changes throughout the second half of the 21st century. For dry periods, RACMO has a very small number of members showing significant changes in both directions, while CRCM and CCLM show marked increases in the number of members with significant positive changes in IAV throughout the 21st century. For the last period 2070–2099, even all members of CCLM show significant increases.

The temporal evolution of IMV (relative to the EM for precipitation-based indicators, Fig. 9) generally supports the direction of changes as seen by the method using the percentage of members with significant changes in IAV. However, when testing for significant changes in the variance between members, hardly any of the changes are significant. CRCM shows significant changes in the majority of years for tas-DJF from 2040 on, for tas-JJA from 2080 on and for pr-JJA from 2060 on. As the IMV is calculated for each year, the plot shows the noise in the IMV per year, which can be large.

Figure 10 shows the change in variability determined from the pooled annual anomalies from the EM for moving 30-year periods from all members. This means, all 30 anomalies from all members are pooled together before calculating the standard deviation (i. e. pooled IAV) and tested for significant changes in the variance with a Brown–Forsythe test. Given the much larger sample size per 30-year period, in contrast to the two former methods, we can now see significant changes in many combinations of indicator and model (Fig. 10). As expected from the previous two methods, internal variability decreases for winter temperature and increases for summer temperature and the number of heat waves. In contrast to the former methods, however, significant changes can be detected earlier. In these cases, the internal variability has already changed significantly in the historical simulations

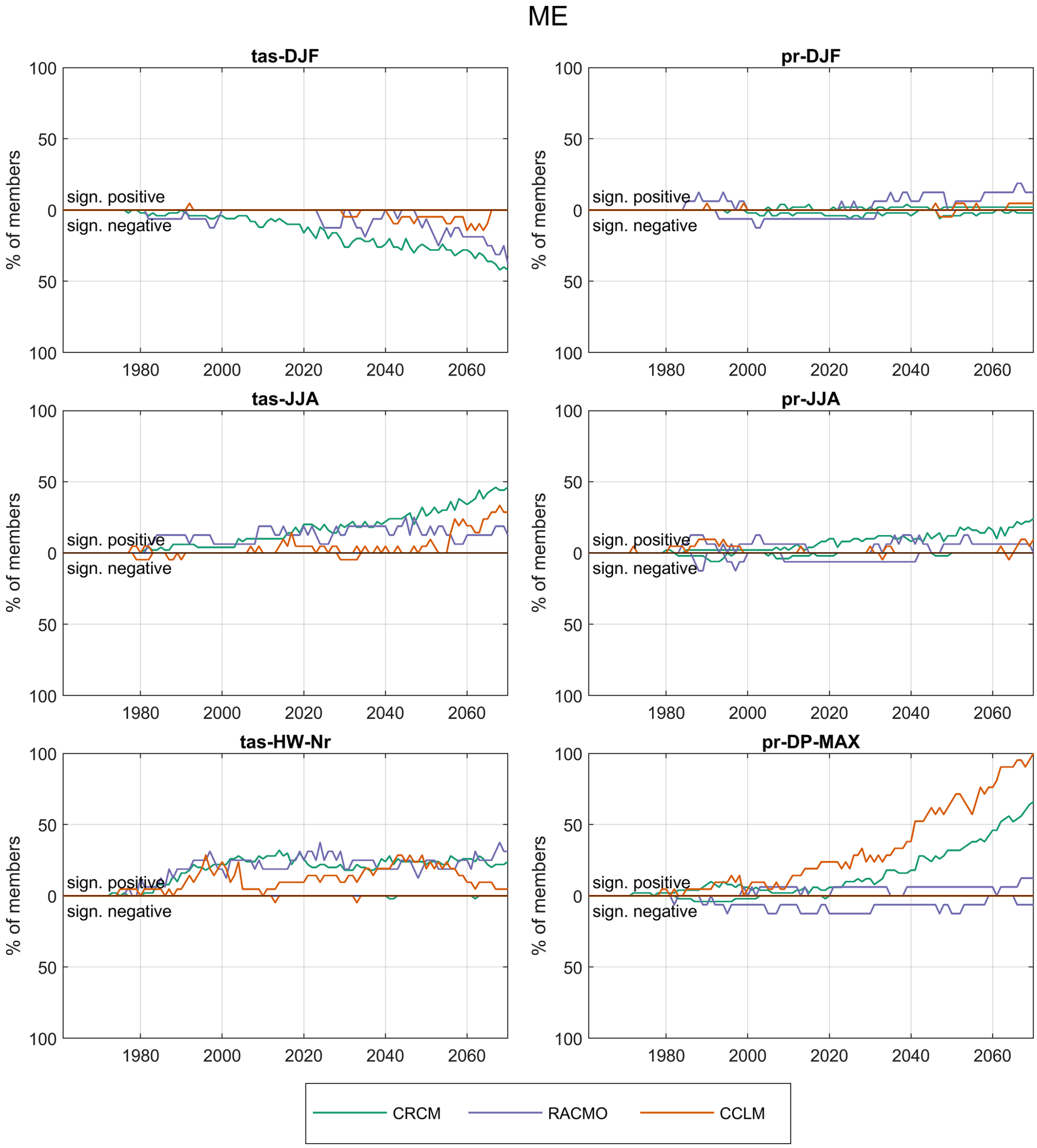

Figure 8. Percentage of members with significantly different variance (Brown–Forsythe test with $\alpha = 0.05$) with respect to the reference period 1961–1990 in mid-Europe. The analysis is based on residuals after removing the EM from each member. The years on the x axis denote to the starting year of moving 30-year periods.

of SMILEs or it changes in the present/near future around 2020. The internal variability in the number of heat waves increases until about 2010–2030, reaches a plateau for about 30–40 years and then decreases again. This behavior can be explained by the forced response of the indicator, which shows strong increases until around 2060, when the number of heat waves stabilizes around 6 (and even decreases afterwards in CRCM, Fig. S2), because the heat waves become so long that their number per year cannot increase anymore. This is especially true for CRCM, where the mean duration of heat waves at the end of the 21st century is much longer than for CCLM and RACMO and about 16 d (not shown), leading to a rough estimate of $6 \times 16 = 96$ heat wave days per year, equal to about 3 months. Since heat waves are defined by the 95th percentile of temperature in the reference period (thus describing extreme conditions), the former extreme heat becomes a regular condition during the summer months at the end of the 21st century in CRCM. For the pooled IAV, both absolute and relative changes in IAV are shown for precipitation-based indicators to demonstrate the effect of the two different approaches. For pr-DJF, CRCM does not show any change in absolute IAV, while this stable behavior in combination with the increase in pr-DJF in the EM leads to a decreasing relative IAV, which is significant

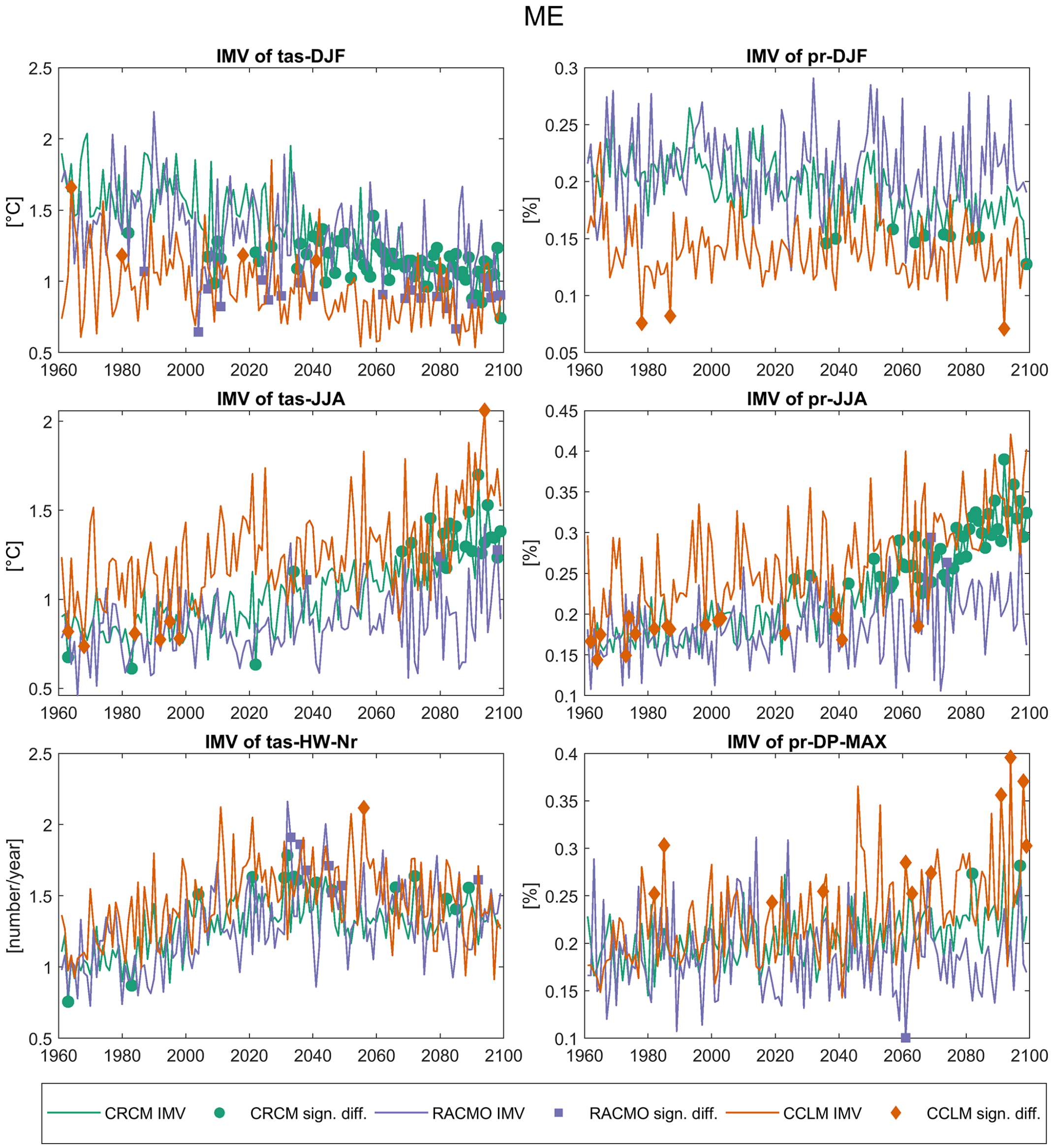

Figure 9. IMV per year sampled on the dimension of the respective ensemble size (50, 21 and 16) for mid-Europe. The analysis is based on residuals after removing the EM from each member. The markers highlight years with a significantly different variance than the reference year 1961. Precipitation-based indicators are shown with their relative anomalies from the ensemble mean (percentage).

from the early 21st century onwards. CCLM and RACMO show increasing variability in absolute terms, but changes are significant for RACMO only, from ∼ 2060 onwards. For both CCLM and RACMO there is no clear change in IAV relative to the change in EM for pr-DJF. Note that while RACMO shows the lowest absolute IAV it shows the highest relative IAV. This originates from the lower EM for winter precipitation in RACMO compared to CRCM and CCLM, which both have quite distinct wet biases (Fig. S2). For summer precipitation, absolute IAV increases according to all models, while EM decreases. Changes in absolute IAV are largest and significant for CRCM and RACMO from ∼ 2000, respectively ∼ 2045 onwards. For CCLM, changes are not significant. Owing to the decreasing EM, increases in relative IAV are significant for all models and significant changes occur earlier in time (∼ 1970 for CCLM, ∼ 1990 for CRCM and ∼ 2040 for RACMO). The changes in EM and IAV in both summer and winter have also been detected by Pendergrass et al. (2017) for CMIP5 and CESM-LE precipitation data in extratropical regions. IAV of pr-DP-MAX does not change according to RACMO, while CRCM and CCLM show distinct increases that go hand in hand with increases in the EM that is also much stronger in these two models than in RACMO (Fig. 4). The changes are significant for relative IAV later in time than for absolute IAV.

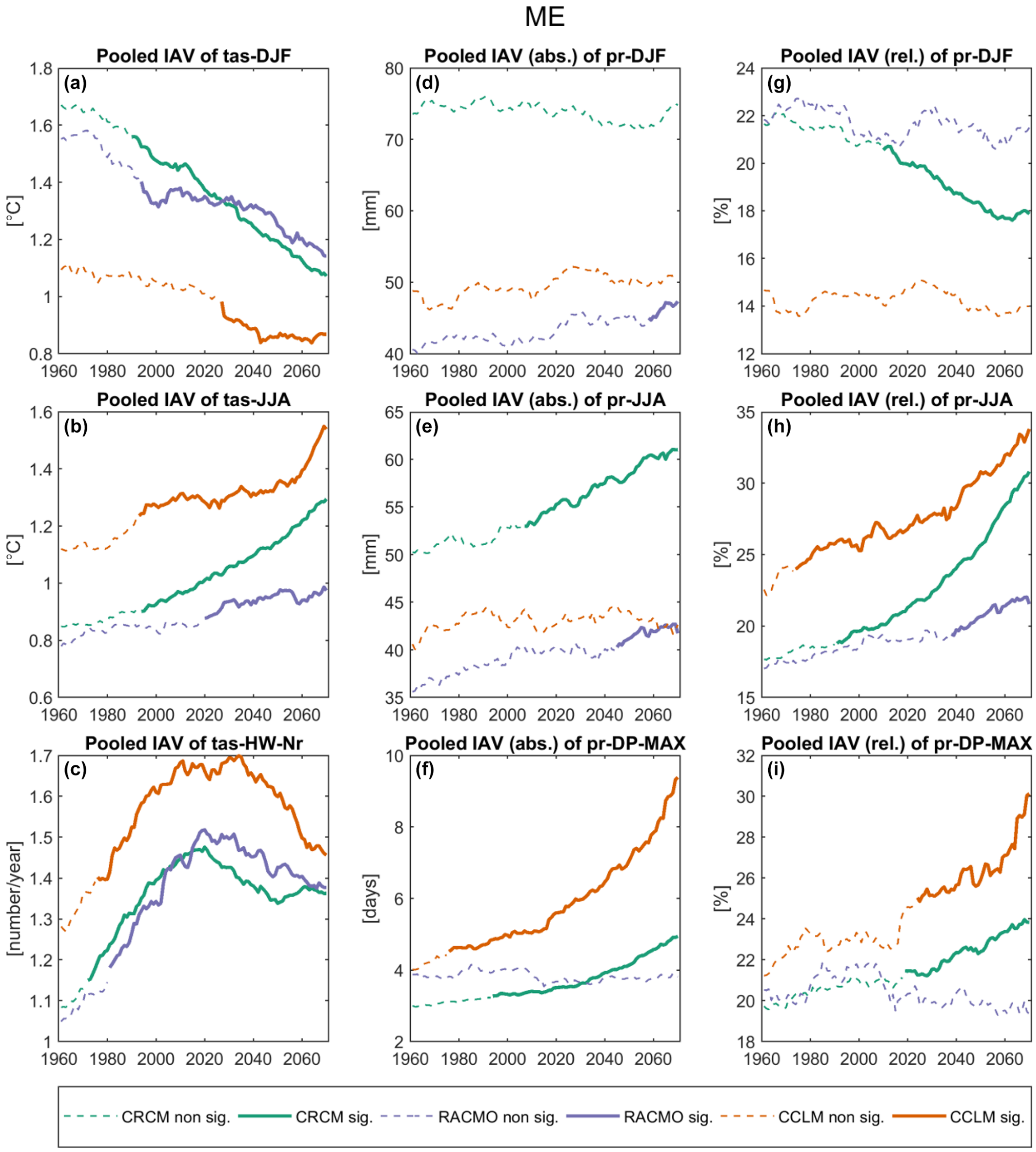

Figure 10. "Pooled IAV" for mid-Europe. The analysis is based on residuals, pooled together from all members, after removing the EM from each member. Temperature-based indicators are shown in absolute terms **(a–c)**. Precipitation-based indicators are shown both in absolute terms **(d–f)** and relative to the ensemble mean **(g–i)**. The change from dashed to solid lines marks the point in time when all following periods show significant changes in variance (BF Test with $\alpha = 0.0.5$).

The abovementioned results are mostly valid for the other regions as well. Differences in the magnitudes of variability and its changes are briefly discussed in the following (see Figs. S8–S10): ME shows higher winter temperature variability than the other three regions, especially than BI. Lower levels of variability compared to the other regions occur over the British Isles for winter temperature and winter precipitation (relative to EM). The Alps show a smaller variability for the number of heat waves than the other regions. The variability of pr-DP-MAX for all three ensembles is similar to ME in AL, while BI and FR hardly show any significant changes. If all regions are considered, RACMO generally has the highest internal variability in winter and the lowest variability in summer for temperature and precipitation (relative to EM), while CCLM has the highest internal variability for summer temperature and precipitation (relative to EM) as well as for heat waves and dry periods (both absolute and relative to EM). Significant changes generally occur similar to ME for winter and summer temperature and precipitation (both absolute and relative to EM). Changes in the number of heat waves are not significant in CCLM in all three regions and in RACMO in AL.

The first method, testing the percentage of members with significant changes in IAV, gives a good overview of the behavior of the members in general. It can, however, only inform about the direction of change in IAV. Additional information on the magnitude of the changes in IAV is needed to get the whole picture. The second method using IMV is in general agreement with the first method when looking at the direction of change. It incorporates the magnitude of internal variability and can show significant changes in the IMV in the same figure. The variations in IMV from year to year are relatively large. Therefore, the BF test results also largely depend on the choice of a reference year to test all other years against. Changes from one year to the next can give very different results. Unstable significance testing is the result. Method 1 is less sensitive to the reference given the overlapping periods. Considering the sample sizes in this study, the best results can be obtained with method 3, where sensitivity can be tested with a much larger sample size. For method 1 it is 30 values per period, for method 2 it is 16/21/50 values per year, and for method 3 it is the product of both 30 years and the ensemble size of the respective ensemble. However, if the sample size was larger for the first two methods, they would probably also result in the detection of significant changes, e.g., when using more members.

To see how the ensemble size impacts the results for the pooled IAV, we reduce the largest ensemble (CRCM with 50 members) to the size of the other ensembles (16 and 21) and repeat the analysis for ME. Even after reducing the ensemble size to only 16 members, the changes at the end of the 21st century in CRCM are still significant for indicators that already showed significant changes for all 50 members. However, the detection of significant changes is only possible at a later time horizon (Fig. S11) for changes in tas-DJF (around 40 years later), relative changes in pr-DJF (20), absolute changes in pr-JJA (20) and relative changes in pr-DP-MAX (15). Tests with a number of ensemble sizes suggest that around 10 members are sufficient to detect the significance of changes and around 20 are sufficient to detect the timing of these significant changes additionally.

4 Discussion

The number of SMILEs available for the quantification of internal variability in this study is still relatively small – we only used three GCM-RCM combinations (to the knowledge of the authors these three ensembles are the only regionally downscaled SMILEs over Europe). More simulations with RCM-SMILEs could help to make results even more robust – especially for winter precipitation and dry periods, where the three ensembles do not agree on the change in variability.

The effect of regional aggregation after the calculation of indicators on the grid level and the potential effects of the original resolution of different data sets on the internal variability seem to be minor for the selected indicators, as seen in the experimental analysis conducted on a subset of the data (Fig. S1). This estimate of sensitivity to differing spatial resolutions might be conservative, however. Nevertheless, the methodology seems to be suitable for the selected indicators of this study.

Methods based on anomalies from the EM are chosen in order to compare results despite different biases in historical and future mean climate states. It can, however, not be ruled out that differences in the variability may originate from mean state biases of the models. CRCM for example shows much higher precipitation sums than the other two ensembles, leading to higher variability in absolute terms. The normalization with the ensemble mean covers these differences in absolute amounts. In the end it largely depends on the definition of variability: is one interested in absolute deviations (mm) or in the fluctuations in relative terms (%)? Results are sensitive to a relative versus an absolute definition or vice versa. The relative approach has the advantage that it allows for a fair comparison of models with different mean precipitation amounts. This is also why a recent publication by Giorgi et al. (2019) gave preference to the relative definition, for example.

The scatterplots of projected changes for seasonal temperature and precipitation (Figs. 2 and 3) show both agreement and dissent, but usually at least two of the three models show similar ranges for one variable. Better agreement might be possible when comparing the data sets not for a fixed period but for periods with the same global warming level in each driving GCM.

We find that the large ensembles analyzed here generally represent observed IAV correctly, but care needs to be taken during the analysis for specific regions and indicators. Cases of many individual members showing both higher and lower variability compared to observational IAV can be found for all ensembles for specific indicators and regions. However, the single observed realization of historical climate makes it difficult to evaluate systematic errors of the ensembles, as the E-OBS distribution is not necessarily representative of the perfectly sampled IAV. It would be interesting to compare large ensembles against an observational large ensemble as proposed by McKinnon and Deser (2018) to better see systematic deficiencies of large ensembles compared to observations.

The results found for changes in IAV are generally in line with existing literature on Europe. We likewise find increasing variability for the summer indicators tas-JJA and pr-JJA (Fischer and Schär, 2009, 2010; Vidale et al., 2007; Yettella et al., 2018; Suarez-Gutierrez et al., 2018) and decreasing variability for the winter indicators tas-DJF and pr-DJF (Bengtsson and Hodges, 2019; Holmes et al., 2016). The summer extreme indicators tas-HW-Nr and pr-DP-MAX also show increased variability in two of the three models, in conjunction with increases in their mean states. Several mechanisms contribute to the changes in all indicators. For changes in the summer temperature IAV, land–atmosphere coupling is becoming more important in central or northern Europe in the future because the transitional zone between dry and wet climates moves northwards from the Mediterranean region, leading to enhanced alternation of dry and wet summer soil moisture (Seneviratne et al., 2006; Fischer et al., 2011). Moreover, stronger warming over land than over the oceans causes the land–ocean temperature gradient in summer to increase. This results in increased variability in thermal advection, which is suggested to play a role in the increase in temperature variability in Europe as well (Holmes et al., 2016). Analysis of observations shows that in the Mediterranean more than half of summer temperature variability can be explained by large-scale atmospheric circulations and sea surface temperatures (Xoplaki et al., 2003). The decrease in winter temperature IAV is suggested to be influenced by changing circulation patterns (Vautard and Yiou, 2009), and a decrease in variability of advected heat due to the decrease in the winter land–ocean temperature gradient (Holmes et al., 2016) and arctic amplification and sea ice loss (Screen, 2014; Sun et al., 2015; Tamarin-Brodsky et al., 2020), even under unchanged circulation variability (Holmes et al., 2016; Tamarin-Brodsky et al., 2020).

The increase in summer precipitation variability that would not be expected under decreasing mean summer precipitation might be caused by a reduction in the number of wet days (>1 mm) that exists in all three ensembles (not shown), as discussed by Räisänen (2002). A reduction in wet days implies an increase in variability since the seasonal precipitation sum becomes more dependent on individual precipitation events.

Land–atmosphere feedback mechanisms are not yet fully understood, and there are still improvements needed in their implementation in earth system models and regional climate models (Vogel et al., 2018). Uncertainties in the future regional development of heat waves and dry periods are thus rather large (Miralles et al., 2019). Nevertheless, increasing frequency, intensity and variability in the number of heat waves as projected by SMILEs using RCP8.5 in this study seem plausible, although the magnitudes can be uncertain. The strong increase in the maximum length of dry periods in two of the models is not necessarily what could be expected. While the length of severe dry periods increases in the future for southern Europe, central and northern Europe do not show any change in the EURO-CORDEX data for dry period length (Jacob et al., 2014). Analysis of precipitation changes shows that both CRCM and CCLM (the RCM itself, not SMILE) are on the dry end of projections for summer precipitation (von Trentini et al., 2019). This might be related to sensitive implementations of land surface modules in these two RCMs. RACMO does not show an increase in the maximum length of dry periods.

Although beyond the scope of this paper, which only analyzed the manifestations of internal model variability in surface variables (tas, pr and associated indicators), there is a need for a better understanding of the mechanisms leading to the model-inherent characteristics of internal variability and why differences between the models appear (e.g., circulation patterns, ocean characteristics).

Using SMILEs for studying changes in IAV allows for much more robust statements on the direction, magnitude and emergence of changes, when using a certain model and RCP scenario. Analyzing the individual members, significant changes in IAV are found in less than half of the members for almost all indicators and ensembles, even at the end of the 21st century (Fig. 8). However, pooling the data of all ensemble members, all three ensembles show significant changes in internal variability of most indicators and often from early in the 21st century onwards (Fig. 10).

Thompson et al. (2015) showed that a statistical model based on a historical period could be as good as a SMILE for predicting future variability of seasonal temperature and precipitation trends up to 2060. The detection of significant increases in internal variability in summer and winter temperature much earlier than 2060 (Fig. 10) challenges this assumption of stable variability. For precipitation (especially absolute changes), however, the changes are often not significant before 2060, confirming the results of Thompson et al. (2015).

5 Conclusions

There is an increasing interest on the part of the scientific community to use single-model initial-condition large ensembles in a wide variety of applications, ranging from deeper levels of understanding of natural climate variability to impact assessments in different fields. The rich data basis which these ensembles provide for the analysis of internal variability is very valuable and enables new insights into this critical part of the climate system. Future changes, in particular, can be better set into context. The effects of dynamical downscaling of GCM large ensembles with regional climate models are not yet sufficiently explored. Further research is needed in this direction to see whether and by how much the internal variability is altered in the RCM simulations of a respective GCM large ensemble. However, downscaling is an important step to make climate simulation information attractive for local adaptation research and impact modellers. The results from this study can be helpful for these research communities to better understand and quantify the role of IAV in the climate system. In particular, increases in variability as seen for summer temperature, relative summer precipitation, heat waves and dry periods in most regions and models can be a huge burden for sectors like agriculture, ecology and hydrology.

The evaluation and comparison of the three RCM-SMILEs in this study gives a first overview of the agreement of SMILEs with observations and among each other. The generally high agreement with observations suggest that the internal variabilities of the RCM-SMILEs at the regional scale are useful approximations of the observed IAV of the climate system in Europe. The direction of changes in internal variability is also mostly the same between the ensembles, suggesting a relatively robust signal. While the "summer indicators" mostly show increasing variability in the future, winter temperature and precipitation show decreasing variability or no change. The change in variability is potentially impact-relevant as it suggests that the most extreme summers and winters may warm more strongly than the corresponding mean.

Despite an increasing number of studies that compare SMILEs (Deser et al., 2014; Martel et al., 2018; Rondeau-Genesse and Braun, 2019; Deser et al., 2020; Lehner et al., 2020), one limitation of many publications using SMILEs is the use of only one model and thereby one estimate of internal variability, leaving it unclear how representative the results are. Although the respective SMILE is usually evaluated against observations in these studies, the uncertainty in future changes in IAV cannot be quantified in the same way as in this study. A further challenge is also the fact that low-frequency variability at decadal and multi-decadal timescales remains uncertain and cannot be rigorously evaluated against observations due to the relatively short observational record and the difficulty of separating forced changes from unforced internal variability in observations.

Overall our results underline the great potential of SMILEs in quantifying the changes in IAV and when they become significant, also at the regional scale.

Data availability. The CRCM5 data are publicly available at https://climex-data.srv.lrz.de/Public/CanESM2_driven_50_members/ (Ouranos, 2020). RACMO data are available upon request directed at erik.van.meijgaard@knmi.nl. CCLM output is available upon request directed at erich.fischer@env.ethz.ch.

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/esd-11-1013-2020-supplement.

Author contributions. FvT designed the concept of the study, performed the analysis and created all figures. EEA and EMF provided the RACMO and CCLM data and helped improve the concept and analysis. FvT led the paper writing with input from all authors.

Competing interests. The authors declare that they have no conflict of interest.

Special issue statement. This article is part of the special issue "Large Ensemble Climate Model Simulations: Exploring Natural Variability, Change Signals and Impacts". It is not associated with a conference.

Acknowledgements. We acknowledge Urs Beyerle, Daniel Lüthi and Sven Kotlarski for their support in setting up the multi-member CCLM experiment. We thank Erik van Meijgaard and Camiel Severijns from KNMI for producing the 16-member KNMI RACMO-EC-EARTH and EC-EARTH ensembles. The CRCM5 was developed by the ESCER centre of Université du Québec à Montréal (UQAM; http://www.escer.uqam.ca, last access: 12 November 2020) in collaboration with Environment and Climate Change Canada. We acknowledge Environment and Climate Change Canada's Canadian Centre for Climate Modelling and Analysis for executing and making available the CanESM2 large ensemble simulations used in this study and the Canadian Sea Ice and Snow Evolution Network for proposing the simulations. The production of ClimEx was funded within the ClimEx project by the Bavarian State Ministry for the Environment and Consumer Protection. Computations with the CRCM5 for the ClimEx project were made on the SuperMUC supercomputer at the Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities. The operation of this supercomputer is funded via the Gauss Centre for Supercomputing (GCS) by the German Federal Ministry of Education and Research and the Bavarian State Ministry of Education, Science and the Arts.

Financial support. This research has been supported by the European Commission, H2020 Research Infrastructures (IMPREX (grant no. GA 641811) and EUCP (grant no. GA 776613)).

Review statement. This paper was edited by Sebastian Milinski and reviewed by three anonymous referees.

References

Aalbers, E. E., Lenderink, G., van Meijgaard, E., and van den Hurk, B. J. J. M.: Local-scale changes in mean and heavy precipitation in Western Europe, climate change or internal variability?, Clim. Dynam., 50, 4745–4766, https://doi.org/10.1007/s00382-017-3901-9, 2018.

Addor, N. and Fischer, E. M.: The influence of natural variability and interpolation errors on bias characterization in RCM simulations, J. Geophys. Res. Atmos., 120, 10180–10195, https://doi.org/10.1002/2014JD022824, 2015.

Bengtsson, L. and Hodges, K. I.: Can an ensemble climate simulation be used to separate climate change signals from internal unforced variability?, Clim. Dynam., 52, 3553–3573, https://doi.org/10.1007/s00382-018-4343-8, 2019.

Brönnimann, S., Rajczak, J., Fischer, E. M., Raible, C. C., Rohrer, M., and Schär, C.: Changing seasonality of moderate and extreme precipitation events in the Alps, Nat. Hazards Earth Syst. Sci., 18, 2047–2056, https://doi.org/10.5194/nhess-18-2047-2018, 2018.

Christensen, J. H. and Christensen, O. B.: A summary of the PRUDENCE model projections of changes in European climate by the end of this century, Climatic Change, 81, 7–30, https://doi.org/10.1007/s10584-006-9210-7, 2007.

Deser, C., Phillips, A. S., Alexander, M. A., and Smoliak, B. V.: Projecting North American Climate over the Next 50 Years: Uncertainty due to Internal Variability, J. Climate, 27, 2271–2296, https://doi.org/10.1175/JCLI-D-13-00451.1, 2014.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., and Ting, M.: Insights from Earth system model initial-condition large ensembles and future prospects, Nat. Clim. Chang., 10, 277–286, https://doi.org/10.1038/s41558-020-0731-2, 2020.

Ferguson, C. R., Pan, M., and Oki, T.: The Effect of Global Warming on Future Water Availability: CMIP5 Synthesis, Water Resour. Res., 54, 7791–7819, https://doi.org/10.1029/2018WR022792, 2018.

Fischer, E. M. and Schär, C.: Future changes in daily summer temperature variability: driving processes and role for temperature extremes, Clim. Dynam., 33, 917–935, https://doi.org/10.1007/s00382-008-0473-8, 2009.

Fischer, E. M. and Schär, C.: Consistent geographical patterns of changes in high-impact European heatwaves, Nat. Geosci., 3, 398–403, https://doi.org/10.1038/ngeo866, 2010.

Fischer, E. M., Lawrence, D. M., and Sanderson, B. M.: Quantifying uncertainties in projections of extremes – a perturbed land surface parameter experiment, Clim. Dynam., 37, 1381–1398, https://doi.org/10.1007/s00382-010-0915-y, 2011.

Fischer, E. M., Rajczak, J., and Schär, C.: Changes in European summer temperature variability revisited, Geophys. Res. Lett., 39, L19702, https://doi.org/10.1029/2012GL052730, 2012.

Fischer, E. M., Beyerle, U., and Knutti, R.: Robust spatially aggregated projections of climate extremes, Nat. Clim. Change, 3, 1033–1038, https://doi.org/10.1038/nclimate2051, 2013.

Giorgi, F.: Dependence of the surface climate interannual variability on spatial scale, Geophys. Res. Lett., 29, 16-1–16-4, https://doi.org/10.1029/2002GL016175, 2002.

Giorgi, F.: Thirty Years of Regional Climate Modeling: Where Are We and Where Are We Going next?, J. Geophys. Res.-Atmos., 124, 5696–5723, https://doi.org/10.1029/2018JD030094, 2019.

Giorgi, F., Bi, X., and Pal, J.: Mean, interannual variability and trends in a regional climate change experiment over Europe. II: climate change scenarios (2071–2100), Clim. Dynam., 23, 839–858, https://doi.org/10.1007/s00382-004-0467-0, 2004.

Giorgi, F., Jones, C., and Asrar, G. R.: Addressing climate information needs at the regional level: the CORDEX framework, WMO Bulletin, 58, 175–183, 2009.

Giorgi, F., Raffaele, F., and Coppola, E.: The response of precipitation characteristics to global warming from climate projections, Earth Syst. Dynam., 10, 73–89, https://doi.org/10.5194/esd-10-73-2019, 2019.

Hawkins, E. and Sutton, R.: The Potential to Narrow Uncertainty in Regional Climate Predictions, B. Am. Meteorol. Soc., 90, 1095–1107, https://doi.org/10.1175/2009BAMS2607.1, 2009.

Hofstra, N., Haylock, M., New, M., and Jones, P. D.: Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature, J. Geophys. Rese.-Atmos., 114, D21101, https://doi.org/10.1029/2009JD011799, 2009.

Holmes, C. R., Woollings, T., Hawkins, E., and de Vries, H.: Robust Future Changes in Temperature Variability under Greenhouse Gas Forcing and the Relationship with Thermal Advection, J. Climate, 29, 2221–2236, https://doi.org/10.1175/JCLI-D-14-00735.1, 2016.

Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O. B., Bouwer, L. M., Braun, A., Colette, A., Déqué, M., Georgievski, G., Georgopoulou, E., Gobiet, A., Menut, L., Nikulin, G., Haensler, A., Hempelmann, N., Jones, C., Keuler, K., Kovats, S., Kröner, N., Kotlarski, S., Kriegsmann, A., Martin, E., van Meijgaard, E., Moseley, C., Pfeifer, S., Preuschmann, S., Radermacher, C., Radtke, K., Rechid, D., Rounsevell, M., Samuelsson, P., Somot, S., Soussana, J.-F., Teichmann, C., Valentini, R., Vautard, R., Weber, B., and Yiou, P.: EURO-CORDEX: New high-resolution climate change projections for European impact research, Reg. Environ. Change, 14, 563–578, https://doi.org/10.1007/s10113-013-0499-2, 2014.

Kendon, E. J., Rowell, D. P., Jones, R. G., and Buonomo, E.: Robustness of Future Changes in Local Precipitation Extremes, J. Climate, 21, 4280–4297, https://doi.org/10.1175/2008JCLI2082.1, 2008.

Kirchmeier-Young, M. C., Zwiers, F. W., and Gillett, N. P.: Attribution of Extreme Events in Arctic Sea Ice Extent, J. Climate, 30, 553–571, https://doi.org/10.1175/JCLI-D-16-0412.1, 2017.

Kotlarski, S., Keuler, K., Christensen, O. B., Colette, A., Déqué, M., Gobiet, A., Goergen, K., Jacob, D., Lüthi, D., van Meijgaard, E., Nikulin, G., Schär, C., Teichmann, C., Vautard, R., Warrach-Sagi, K., and Wulfmeyer, V.: Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble, Geosci. Model Dev., 7, 1297–1333, https://doi.org/10.5194/gmd-7-1297-2014, 2014.

Kumar, D. and Ganguly, A. R.: Intercomparison of model response and internal variability across climate model ensembles, Clim. Dynam., 51, 207–219, https://doi.org/10.1007/s00382-017-3914-4, 2018.

Leduc, M., Mailhot, A., Frigon, A., Martel, J.-L., Ludwig, R., Brietzke, G. B., Giguére, M., Brissette, F., Turcotte, R., Braun, M., and Scinocca, J.: 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. Meteorol. Clim., https://doi.org/10.1175/JAMC-D-18-0021.1, 2019.

Lehner, F., Deser, C., and Sanderson, B. M.: Future risk of record-breaking summer temperatures and its mitigation, Climatic Change, 146, 363–375, https://doi.org/10.1007/s10584-016-1616-2, 2018.

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

Lenderink, G.: Exploring metrics of extreme daily precipitation in a large ensemble of regional climate model simulations, Clim. Res., 44, 151–166, https://doi.org/10.3354/cr00946, 2010.

Lorenz, P. and Jacob, D.: Validation of temperature trends in the ENSEMBLES regional climate model runs driven by ERA40, Clim. Res., 44, 167–177, https://doi.org/10.3354/cr00973, 2010.

Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., Kröger, J., Takano, Y., Ghosh, R., Hedemann, C., Li, C., Li, H., Manzini, E., Notz, D., Putrasahan, D., Boysen, L., Claussen, M., Ilyina, T., Olonscheck, D., Raddatz, T., Stevens, B., and Marotzke, J.: The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability, J. Adv. Model. Earth Sy., 11, 2050–2069, https://doi.org/10.1029/2019MS001639, 2019.

Martel, J.-L., Mailhot, A., Brissette, F., and Caya, D.: Role of Natural Climate Variability in the Detection of Anthropogenic Climate Change Signal for Mean and Extreme Precipitation at Local and Regional Scales, J. Climate, 31, 4241–4263, https://doi.org/10.1175/JCLI-D-17-0282.1, 2018.

McKinnon, K. A. and Deser, C.: Internal Variability and Regional Climate Trends in an Observational Large Ensemble, J. Climate, 31, 6783–6802, https://doi.org/10.1175/JCLI-D-17-0901.1, 2018.

Miralles, D. G., Gentine, P., Seneviratne, S. I., and Teuling, A. J.: Land-atmospheric feedbacks during droughts and heatwaves: state of the science and current challenges, Ann. NY Acad. Sci., 1436, 19–35, https://doi.org/10.1111/nyas.13912, 2019.

Nikiéma, O., Laprise, R., and Dugas, B.: Energetics of transient-eddy and inter-member variabilities in global and regional climate model simulations, Clim. Dynam., 51, 249–268, https://doi.org/10.1007/s00382-017-3918-0, 2018.

Ouranos: CRCM5-LE, ClimEx, available at: https://climex-data.srv.lrz.de/Public/CanESM2_driven_50_members/, last access: 16 November 2020.

Pendergrass, A. G., Knutti, R., Lehner, F., Deser, C., and Sanderson, B. M.: Precipitation variability increases in a warmer climate, Sci. Rep., 7, 17966, https://doi.org/10.1038/s41598-017-17966-y, 2017.

Räisänen, J.: CO_2-Induced Changes in Interannual Temperature and Precipitation Variability in 19 CMIP2 Experiments, J. Climate, 15, 2395–2411, https://doi.org/10.1175/1520-0442(2002)015<2395:CICIIT>2.0.CO;2, 2002.

Rondeau-Genesse, G. and Braun, M.: Impact of internal variability on climate change for the upcoming decades: analysis of the CanESM2-LE and CESM-LE large ensembles, Climatic Change, 156, 299–314, https://doi.org/10.1007/s10584-019-02550-2, 2019.

Screen, J. A.: Arctic amplification decreases temperature variance in northern mid- to high-latitudes, Nat. Clim. Change, 4, 577–582, https://doi.org/10.1038/nclimate2268, 2014.

Seneviratne, S. I., Lüthi, D., Litschi, M., and Schär, C.: Land-atmosphere coupling and climate change in Europe, Nature, 443, 205–209, https://doi.org/10.1038/nature05095, 2006.

Sørland, S. L., Schär, C., Lüthi, D., and Kjellström, E.: Bias patterns and climate change signals in GCM-RCM model chains, Environ. Res. Lett., 13, 74017, https://doi.org/10.1088/1748-9326/aacc77, 2018.

Suarez-Gutierrez, L., Li, C., Müller, W. A., and Marotzke, J.: Internal variability in European summer temperatures at 1.5 °C and 2 °C of global warming, Environ. Res. Lett., 13, 64026, https://doi.org/10.1088/1748-9326/aaba58, 2018.

Sun, L., Deser, C., and Tomas, R. A.: Mechanisms of Stratospheric and Tropospheric Circulation Response to Projected Arctic Sea Ice Loss, J. Climate, 28, 7824–7845, https://doi.org/10.1175/JCLI-D-15-0169.1, 2015.

Tamarin-Brodsky, T., Hodges, K., Hoskins, B. J., and Shepherd, T. G.: Changes in Northern Hemisphere temperature variability shaped by regional warming patterns, Nat. Geosci., 13, 414–421, https://doi.org/10.1038/s41561-020-0576-3, 2020.

Thompson, D. W. J., Barnes, E. A., Deser, C., Foust, W. E., and Phillips, A. S.: Quantifying the Role of Internal Climate Variability in Future Climate Trends, J. Climate, 28, 6443–6456, https://doi.org/10.1175/JCLI-D-14-00830.1, 2015.

Torma, C., Giorgi, F., and Coppola, E.: Added value of regional climate modeling over areas characterized by complex terrain-Precipitation over the Alps, J. Geophys. Res.-Atmos., 120, 3957–3972, https://doi.org/10.1002/2014JD022781, 2015.

Vautard, R. and Yiou, P.: Control of recent European surface climate change by atmospheric flow, Geophys. Res. Lett., 36, 231, L22702, https://doi.org/10.1029/2009GL040480, 2009.

Vidale, P. L., Lüthi, D., Wegmann, R., and Schär, C.: European summer climate variability in a heterogeneous multi-model ensemble, Climatic Change, 81, 209–232, https://doi.org/10.1007/s10584-006-9218-z, 2007.

Vogel, M. M., Zscheischler, J., and Seneviratne, S. I.: 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, 2018.

von Trentini, F., Leduc, M., and Ludwig, R.: Assessing natural variability in RCM signals: comparison of a multi model EURO-CORDEX ensemble with a 50-member single model large ensemble, Clim. Dynam., 53, 1963–1979, https://doi.org/10.1007/s00382-019-04755-8, 2019.

Wang, L., Deng, A., and Huang, R.: Wintertime internal climate variability over Eurasia in the CESM large ensemble, Clim. Dynam., 52, 6735–6748, https://doi.org/10.1007/s00382-018-4542-3, 2019.

Xoplaki, E., González-Rouco, J. F., Luterbacher, J., and Wanner, H.: Mediterranean summer air temperature variability and its connection to the large-scale atmospheric circulation and SSTs, Clim. Dynam., 20, 723–739, https://doi.org/10.1007/s00382-003-0304-x, 2003.

Yettella, V., Weiss, J. B., Kay, J. E., and Pendergrass, A. G.: An Ensemble Covariance Framework for Quantifying Forced Climate Variability and Its Time of Emergence, J. Climate, 31, 4117–4133, https://doi.org/10.1175/JCLI-D-17-0719.1, 2018.