Climate Dynamics (2019) 52:3575–3598
https://doi.org/10.1007/s00382-018-4344-7

Classifications of winter atmospheric circulation patterns: validation of CMIP5 GCMs over Europe and the North Atlantic

Jan Stryhal[1,2] · **Radan Huth**[1,2]

Received: 11 August 2017 / Accepted: 30 June 2018 / Published online: 7 July 2018

Abstract
Winter atmospheric circulation over the Euro-Atlantic domain and three subdomains (British Isles, Central Europe, and Eastern Mediterranean) is validated in outputs of historical runs of 32 global climate models (GCMs) from phase 5 of the Coupled Model Intercomparison Project (CMIP5). Eight automated classifications of daily SLP patterns from five reanalysis datasets are produced for each domain in order to analyse the effect of the choices of methods and reference data on results. The results show that the ranking of GCMs fundamentally depends on which classification is used; therefore, only parallel usage of multiple classifications can provide robust rankings of models. Considering all eight classifications, three models (HadGEM2-CC, MIROC4h, and CNRM-CM5) are among the best in simulating the frequency of circulation types (CTs) over all four domains. Regardless the domain, the bias in CT frequency of the worst GCMs is larger than 50% of the frequency in the reference reanalysis dataset. Conversely, the best GCM for each domain differs from the reference reanalysis by about 10–20%, which is nearly the same result as found for the NOAA-CIRES Twentieth Century Reanalysis (version 2). The persistence of circulation is simulated better than the frequency with errors rarely exceeding 15%. The GCMs overestimate the frequency of westerly circulation over all domains (by about 7% over the British Isles, 21% over Central Europe, and almost 70% over the Eastern Mediterranean) and also cyclonic CTs, while easterly and anticyclonic CTs are typically underestimated by 30–40%.

Keywords Global climate models · Validation · Atmospheric circulation · Circulation classifications

1 Introduction

Global climate models (GCMs) have become an invaluable tool for studying the climate system of the Earth. Validation of these models against observation-based datasets is of utmost importance in order to assess the reliability of the models as well as to provide the developers with a feedback that would help them further improve their models. One of the simulated climatic features that have drawn particular attention of researchers is large-scale atmospheric circulation.

The importance of the ability of climate models to simulate atmospheric circulation and not only thermodynamic and moisture variables was repeatedly discussed (e.g. Boer et al. 1992; McKendry et al. 1995; Hall 2014; Shepherd 2014; Kröner et al. 2017). For the Euro-Atlantic domain, this ability seems particularly important owing to the synoptic link between the large-scale circulation and regional/local near-surface climatic (meteorological) elements, which is especially tight during winter (see e.g. van Ulden and van Oldenborgh 2006; Beck et al. 2007; Plavcová and Kyselý 2013; Broderick and Fealy 2015; Belleflamme et al. 2015; Cahynová and Huth 2016). Consequently, errors in simulation of circulation properties (such as the strength, direction, vorticity, and persistence of flow) and/or of the synoptic link markedly limit the applicability of model output in both statistical and dynamical downscaling, since statistical models assume this link is simulated correctly and regional climate models have only limited ability to improve on the biases inherited from their driving data (van Ulden and van Oldenborgh 2006; Plavcová and Kyselý 2012).

✉ Jan Stryhal
stryhal@ufa.cas.cz

[1] Department of Physical Geography and Geoecology, Faculty of Science, Charles University, Albertov 6, 12843 Prague 2, Czech Republic

[2] Institute of Atmospheric Physics, Czech Academy of Sciences, Boční II, 14131 Prague 4, Czech Republic

Several approaches have been used to characterize atmospheric circulation both in reality (station data) or quasi-reality (atmospheric reanalyses) and in model simulations and projections. One of the most widely used approaches is classifications of atmospheric circulation patterns (circulation classifications for short). They can be seen as a tool that describes the entire variety of atmospheric circulation by defining a catalogue of a few circulation types (CTs) and subsequently classifying circulation patterns with one of these CTs. The classified circulation patterns are usually instantaneous or daily mean sea level pressure (SLP) or geopotential height (GPH) patterns defined typically on regional to continental spatial scales.

Many statistical methods have been used to obtain circulation classifications; the topic was reviewed in detail by Huth et al. (2008). To evaluate circulation in model output, the most widely used approaches have been leader (threshold-based) algorithms (Crane and Barry 1988; McKendry et al. 1995; Lapp et al. 2002; Schoof and Pryor 2006; Belleflamme et al. 2013, 2015), principal component analysis (Huth 1997, 2000), cluster analysis (McKendry et al. 2006; Rust et al. 2010; Pastor and Casado 2012; Cattiaux et al. 2013b; Perez et al. 2014; Rohrer et al. 2017), the neural-network algorithm of self-organizing maps (Cassano et al. 2006; Lynch et al. 2006; Finnis et al. 2009a, b; Gibson et al. 2016), as well as approaches that predefine CTs—and thus constitute hybrids between (the above listed) automated classification methods and traditional manual synoptic catalogues—, for example Jenkinson-Collison's method (Demuzere et al. 2009; Lorenzo et al. 2011; Rohrer et al. 2017).

Over the last three decades, research into circulation types in model simulations has developed from pioneering studies focused on one model and various methodological aspects of the analysis to complex examinations of multi-model ensembles, in response to development in statistics, data processing, and computing and to production of a large amount of GCM outputs under phases three (Meehl et al. 2007) and five (Taylor et al. 2012) of the Coupled Model Intercomparison Project (CMIP). Nevertheless, two issues have so far been rather marginalized in this kind of studies, namely how is the result affected by the choice of a reference dataset reanalysis—which represents the quasi-reality against which models are validated—and a classification method.

So far, most studies have arbitrarily chosen either ERA-40 or NCEP-1 (see Sect. 2.2 for more information on reanalyses) as their reference dataset. A few studies utilised more than one reanalysis (Rust et al. 2010; Belleflamme et al. 2013, 2015; Perez et al. 2014; Gibson et al. 2016); however, the goal of all of these studies was evaluation of models and only little attention was paid to the intercomparison of reanalyses. Recently, Stryhal and Huth (2017) showed that classifications can considerably vary in different reanalyses even over regions where abundant observations were assimilated into reanalyses, such as Europe, and that the choice of reanalysis can have a substantial effect on errors of simulated CT frequencies and, consequently, rankings of GCMs based on these errors. Therefore, they suggested that multiple reanalyses be used in future validation studies. We aim to address this issue by validating CMIP5 GCMs against an ensemble of five reanalyses—such that the observation uncertainty is accounted for and the magnitudes of inter-reanalysis differences and GCM biases can be compared.

Furthermore, classification methods have been used rather arbitrarily as well; only a handful of the studies listed above—namely those by Rust et al. (2010), Pastor and Casado (2012), Belleflamme et al. (2013), Rohrer et al. (2017), and Stryhal and Huth (2017, 2018)—used more (typically two) classifications to compare CTs in multiple datasets. This is somewhat striking since it has been shown nearly two decades ago that parallel examination of multiple classifications helps eliminate subjectivity of the methodology and provides more reliable results if one compares circulation in multiple datasets (Huth 2000). Furthermore, Rust et al. (2010) showed that the magnitude of differences between two reanalyses is sensitive to how CTs are defined. In their validation study, Belleflamme et al. (2013) concluded that also the similarity measure (e.g. Euclidean distance or pattern correlation) used to classify daily patterns with CTs markedly affects the results as it highlights qualitatively different kinds of GCM errors. Stryhal and Huth (2017) showed that differences between reanalyses project into various classifications with varying intensity, causing the estimation of significance of these differences to be very sensitive to subjective methodological choices.

Recently, ample evidence has been accumulated that relying on a single classification provides one with only an incomplete picture of reality and, consequently, puts one at risk of misinterpreting results. This risk was shown to relate to various meteorological and environmental variables and phenomena including temperature (Huth 2010; Ustrnul et al. 2010; Broderick and Fealy 2015; Beck et al. 2016; Huth et al. 2016), precipitation (Casado et al. 2010; Lupikasza 2010; Schiemann and Frei 2010; Tveito 2010; Broderick and Fealy 2015; Beck et al. 2016; Casado and Pastor 2016; Huth et al. 2016), trends in meteorological variables (Cahynová and Huth 2010, 2016) and circulation (Belleflamme et al. 2013; Kučerová et al. 2017), droughts (Fleig et al. 2010; Beck et al. 2015), air quality (Stefan et al. 2010; Beck et al. 2014; Valverde et al. 2015), surface ozone (Demuzere et al. 2011), wild fires (Kassomenos 2010), landslides (Wood et al. 2016), and phenological phases (Palm et al. 2017). Additionally, it was proved beyond doubt that none of the statistical approaches to classification is superior and that which method is the best depends on many factors including (but not limited to) the objective of a study, the classified variable and the target environmental or meteorological

variable, number of CTs, and character of circulation. In the present paper, this issue is addressed by repeating the validation by multiple distinct classification methods—such that a robust estimation of GCM errors is obtained, misinterpretation of results is avoided, and the sensitivity of validation to the choice of methods is quantified.

The goal of the paper is to validate daily mean SLP patterns simulated by an ensemble of 32 CMIP5 GCMs for the historical period of 1961–2000 against reality represented by five atmospheric reanalyses, with emphasis on the sensitivity of the results to the choice of classification method. The study is carried out for winter—a season in which the link between circulation and surface climate is strongest over the region and, therefore, for which the ability of models to simulate circulation correctly is of utmost importance. Four spatial domains are analysed: the continental-scale Euro-Atlantic domain as well as three smaller domains—the British Isles, Central Europe, and the Eastern Mediterranean.

2 Data and methods

2.1 Classification methods

The study can be seen as a follow-up to the European Cooperation in Science and Technology Action 733 (COST 733) "Harmonisation and Applications of Weather Type Classifications for European regions"—see editorials of special issues by Huth et al. (2010) and Tveito and Huth (2016) where results of COST 733 are summarized—which aimed at creating a database of classification methods and their applications in synoptic climatology. Additionally, a software package including 33 methods was coded and made freely available online (http://cost733.geo.uni-augsburg.de/cost733wiki). From this database, we use eight methods (see Table 1), which together represent all main approaches to classifying circulation patterns. A brief description of the methods follows; for more details refer to Philipp et al. (2010, 2016).

First, hybrid methods used here are Grosswettertypes (GWT) and two algorithms of Jenkinson–Collison (JCT). All three algorithms define ten CTs based on pre-set thresholds of several circulation indices, such as vorticity and flow direction. The two JCT algorithms differ only in the selection of grid points from which the indices are calculated: JCT1 (JCT2) selects points from the centre (across the whole) of a region. Eight of the ten CTs are directional—one for advection from each directional octant [further referred to as west (W), northwest (NW), north (N), northeast (NE), east (E), southeast (SE), south (S), and southwest (SW)]—one is purely cyclonic (C), and one purely anticyclonic (AC). Second, Lund's method is one of leader-based algorithms; by comparing pattern-to-pattern correlations, it finds key patterns that well represent relatively large parts of the phase space. Third, methods based on principal component analysis (PCA) are represented by PCA with the input data matrix in the T-mode [i.e. grid points correspond to columns of the data matrix and time realizations (days) to its rows], followed by oblique rotation of principal components (PCs). The scores of the rotated PCs represent CTs and their loadings (which form time series) are used to assign patterns to classes. Last, three algorithms of non-hierarchical CA are used (*k*-means, *k*-medoids, and SANDRA). These three methods—also called optimisation methods—incorporate steps that help find a solution closer to the optimal partitioning (that is the one with minimum within-type variance).

Every automated method requires a few methodological choices to be made before the classification is run. Regarding the number of CTs and the spatial extent of geographical domains, we follow the choices made by the COST733 action. Nine CTs are defined for each classification, with the exception of hybrid methods, for which

Table 1 List of classification methods used in the study

Method abbreviation	Method name	No of CTs	Method group	Projection method	References
GWT	Grosswettertypes	10	Hybrid	None	Beck et al. (2007)
JCT1, JCT2	Jenkinson–Collison	10			James (2006)
LND	Lund	9	Leader algorithm	Highest pattern-to-CT-centroid correlation	Lund (1963)
PCT	T-mode PCA obliquely rotated	9	Principal component analysis	Highest pattern-to-PC-score correlation	Huth (1996)
CKM	*k*-means (differing start partitions)	9	Cluster analysis	Smallest pattern-to-CT-centroid Euclidean distance	Enke and Spekat (1997)
SAN	Simulated annealing and diversified randomisation (SANDRA)	9			Philipp et al. (2007)
KMD	*k*-medoids	9			Kaufman and Rousseeuw (1990)

ten are more feasible as described above. Classifications are performed over four domains, one covering the whole of Europe, others covering the British Isles, Central Europe, and the Eastern Mediterranean (see Fig. 1), in order that the skill of GCMs and the usability of methods are assessed across regions with different climate conditions and circulation. For each dataset and each domain, eight classification methods produce 75 ($3\times10+5\times9$; see Table 1) CTs together. Since the software requires the patterns to have an identical spatial resolution, all datasets were interpolated onto the longitude-latitude grid of 3° per 2° over the large domain and 1° per 1° over the small domains; original resolution of the datasets is included in Tables 2 and 3.

2.2 Reanalyses and GCM simulations

The analyses are carried out for 40 climatological winters (Dec 1960 to Feb 2000). 29 February was excluded for simplicity. For this period, daily mean SLP maps produced by historical runs of 32 CMIP5 GCMs were accessed from the Program for Climate Model Diagnosis and Intercomparison (http://www-pcmdi.llnl.gov) and the Earth System Grid [https://www.earthsystemgrid.org, models CCSM4 and CESM1(CAM5)] databases. Only one ensemble member was chosen for each GCM, preferably r1i1p1. If r1i1p1 was not available, models are represented by another member: r3i1p1 (IPSL-CM5A-MR), r5i1p1 (HadGEM2-ES), r6i1p1 (IPSL-CM5A-LR), r1i2p1 (CCSM4, CSIRO-Mk3L-1-2). GCMs were validated against five reanalysis datasets that fully cover the surveyed period and domains. The lists of GCMs and reanalyses used in the paper are shown in Tables 2 and 3, respectively.

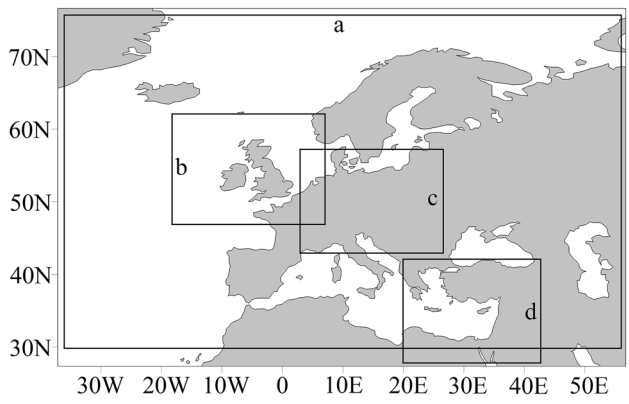


Fig. 1 Spatial extent of domains analysed in the study: **a** Europe and the North Atlantic (domain D00 according to COST733), **b** British Isles (D04), **c** Central Europe (D07), and **d** Eastern Mediterranean (D11)

2.3 Circulation statistics and comparison of datasets

First, using each method, a classification is performed for all (18,000) available reanalysed daily patterns (5 reanalyses × 40 winters × 90 days). The subsequent procedure differs for hybrid algorithms from the rest of methods. Since both GWT and JCT predefine the catalogue of CTs, classifications in reanalyses can be directly compared with classifications in GCM simulations (115,200 patterns), and CT frequency and persistence easily calculated separately for each reanalysis and each model.

For the remaining methods, the shape of CTs [note that by the shape of a CT one understands the composite (centroid) pattern computed as the mean of all patterns classified with the CT (Philipp et al. 2016)] is a result of the classification and, thus, differs for classifications in reanalyses and in models. Therefore, projection is used in order to obtain comparable classifications: first, CTs are defined in combined output of all reanalyses, CT centroids are computed for all reanalyses together (reanalysis centroids), and the circulation statistics are calculated for each reanalysis separately. Second, the centroids are projected onto all simulated patterns, resulting in each pattern being classified with (assigned to) the most similar CT. The projection method (Table 1) is specific for each classification method and is identical to the approach that the particular classification method uses to classify patterns with CTs (e.g. Euclidean distance for *k*-means and pattern correlation for Lund). Last, the circulation statistics are calculated for every model, and centroids for the GCM ensemble (GCM centroids) are created. Additionally to biases in CT frequency and persistence, also biases in the centroids are assessed by means of three metrics: pattern correlation, Euclidean distance, and mean horizontal SLP gradient between the two grid points with the highest and lowest SLP.

To verify the results of projection, the whole process was repeated also in the opposite direction; that is, CTs were first defined on all simulated patterns and, subsequently, projected onto reanalyses. Nevertheless, running classifications on all GCMs together was computationally unexecutable for LND and KMD; therefore, only PCT, CKM, and SAN were verified this way. Ideally, GCM biases should be approximately the same for both directions of projection. Indeed, the results were almost identical in most cases; therefore, only the results for projection from reanalyses to GCMs are presented. The projections from GCMs onto reanalyses are discussed only if they considerably differ from their counterparts.

Additionally to analysing each classification separately, some analyses are carried out on the set of all 75 CTs regardless of the classification method. It has to be stressed that data are not classified to 75 classes but by eight methods into

Table 2 List of GCMs used in the study

No	Model acronym	Model resolution (LON×LAT)	Modelling centre or group
1	BCC_CSM1.1(m)	1.1°×1.1°	Beijing Climate Centre, China Meteorological Administration
2	CanESM2	2.8°×2.8°	Canadian Centre for Climate Modelling and Analysis
3	CCSM4	1.3°×0.9°	National Centre for Atmospheric Research
4	CESM1(CAM5)	1.3°×0.9°	Community Earth System Model Contributors
5	CMCC-CESM	3.8°×3.8°	Centro Euro-Mediterraneo per I Cambiamenti Climatici
6	CMCC-CM	0.8°×0.8°	
7	CMCC-CMS	1.9°×1.9°	
8	CNRM-CM5	1.4°×1.4°	Centre National de Recherches Météorologiques/Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique
9	CSIRO-Mk3L-1-2	5.6°×3.2°	Commonwealth Scientific and Industrial Research Organisation in collaboration with the Queensland Climate Change Centre of Excellence
10	EC-EARTH	1.1°×1.1°	EC-EARTH consortium
11	FGOALS-g2	2.8°×2.8°	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences and CESS, Tsinghua University
12	GFDL_CM3	2.5°×2.0°	NOAA Geophysical Fluid Dynamics Laboratory
13	GFDL-ESM2G	2.5°×2.0°	
14	GFDL-ESM2M	2.5°×2.0°	
15	HadCM3	3.8°×1.9°	Met Office Hadley Centre
16	HadGEM2-AO	1.9°×1.3°	National Institute of Meteorological Research /Korea Meteorological Administration
17	HadGEM2-CC	1.9°×1.3°	Met Office Hadley Centre
18	HadGEM2-ES	1.9°×1.3°	
19	INM-CM4.0	2.0°×1.5°	Institute of Numerical Mathematics
20	IPSL-CM5A-LR	3.8°×1.9°	Institut Pierre-Simon Laplace
21	IPSL-CM5A-MR	2.5°×1.3°	
22	IPSL-CM5B-LR	3.8°×1.9°	
23	MIROC4h	0.6°×0.6°	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology
24	MIROC5	1.4°×1.4°	
25	MIROC-ESM	2.8°×2.8°	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo) and National Institute for Environmental Studies
26	MIROC-ESM-CHEM	2.8°×2.8°	
27	MPI-ESM-LR	1.9°×1.9°	Max Planck Institute for Meteorology
28	MPI-ESM-MR	1.9°×1.9°	
29	MPI-ESM-P	1.9°×1.9°	
30	MRI-CGCM3	1.1° × 1.1°	Meteorological Research Institute
31	MRI-ESM1	1.1°×1.1°	
32	NorESM1-M	2.5°×1.9°	Norwegian Climate Centre

Table 3 List of atmospheric reanalyses used in the study

Acronym	Institute	Resolution (LON×LAT)	References
ERA-40	European Centre for Medium-Range Weather Forecasts	1.125°×1.125°	Uppala et al. (2005)
NCEP-1	National Centers for Environmental Prediction(NCEP)-National Center for Atmospheric Research (NCAR)	2.5°×2.5°	Kalnay et al. (1996)
JRA-55	Japan Meteorological Agency	0.5625°×0.5625°	Kobayashi et al. (2015)
20CRv2	NOAA/Earth System Research Laboratory, University of Colorado CIRES Climate Diagnostics Center	2°×2°	Compo et al. (2011)
ERA-20C	European Centre for Medium-Range Weather Forecasts	2°×2°	Poli et al. (2016)

9 or 10 classes. Since the total relative frequency is 100% for each classification, the relative frequency corresponding to all eight methods, that is, to all 75 CTs, is 800%. This may seem counterintuitive, but it should be noted that every single circulation pattern is classified with eight CTs (one for each method)—hence the total of 800% is quite natural.

3 Results and discussion

In this chapter, main results are presented and discussed. The first four sections focus on general issues concerning the validation: In Sect. 3.1, the reference dataset is selected with which reanalyses and GCMs are compared; in Sect. 3.2, GCMs are ranked according to their performance in terms of frequency and persistence of CTs; Sect. 3.3 describes how GCM biases in CT frequency (and GCM rankings) depend on the choice of classification method; and in Sect. 3.4, the relation between biases in CT frequency and mean SLP is analysed and discussed. Section 3.5 presents the most important biases of GCMs in the four regions and discusses them in more detail. Last, Sect. 3.6 addresses main limitations of the study.

3.1 Observational uncertainty and selection of the reference dataset

Since no study has so far validated the large-scale circulation over Europe in reanalyses against independent observations, it is not known which reanalysis is the best. Therefore, rather than arbitrarily choosing one of the reanalyses as a reference, all datasets (both reanalyses and GCMs) are compared with a "median" reanalysis. The median reanalysis is defined for each variable (e.g. frequency of a CT) as the median of the corresponding values in the five reanalyses.

In Fig. 2, the difference of CT frequency in individual reanalyses from the median reanalysis is shown for each domain; furthermore, these differences are compared to biases of GCMs, which are defined as the median of absolute values of errors of the frequency of 75 CTs. The results show that—with the exception of 20CRv2 and the Eastern Mediterranean—reanalyses differ from the reference value, and therefore from each other, relatively little. Furthermore, ERA-40 is in all four cases closest to the reanalysis median. Although this does not necessarily mean that ERA-40 is the best reanalysis, it does mean that using ERA-40 alone to validate the models would lead, relative to selecting any of the other reanalyses, to results most similar to the considerably more laborious approach used here. It should be noted that although the *median* deviation is small, deviations of some CTs are not to be neglected. For instance, over the British Isles, the second most frequent CT in the PCT classification (SW directional CT) is 14% more frequent in NCEP-1 relative to ERA-40. Readers are referred to Stryhal and Huth (2017) for more information on the effect of the choice of classification on differences between reanalyses; the effect of classifications on GCM biases will be analysed in detail later in the text.

One might argue that the anomalous behaviour of 20CRv2 means that one should exclude this reanalysis from validation. We include 20CRv2 in order that one can see how it performs relative to good GCMs; moreover, since medians are used instead of means, including one outlying value does not have a marked impact on results.

3.2 Ranking of models

The median absolute errors shown in Fig. 2 can be used to rank GCMs. The median error of the best performing model for each domain is between 10 and 20%, which is, interestingly, approximately as far from the median reanalysis as the outlying reanalysis, 20CRv2. On the contrary, for all domains, several worst models differ from the median reanalysis by more than 50% (i.e. the frequency of at least half of the CTs is either less than 50% or more than 150% of the median reanalysis). A further investigation of GCM errors in CT frequency revealed that what particularly discriminates good models from the bad ones is the ability to simulate the occurrence of frequent CTs (those with the relative frequency of 15% and more), while the skill to simulate less frequent CTs is very similar for most models (not shown). Additionally, GCMs are considerably better at simulating persistence than frequency: median errors of persistence of most GCMs do not exceed 15% (Fig. 3).

Figure 4 compares GCM rankings for all domains based on the biases of frequency and persistence. Notably, there are not one or a few GCMs outperforming the rest of models in simulating CTs for all domains. Based on frequency, only three models rank among the best ten models for all four domains: HadGEM2-CC, MIROC4h, and CNRM-CM5. Moreover, a different model ranks first for each domain: HadGEM2-CC for the Euro-Atlantic domain, GFDL-ESM2G for the British Isles, CMCC-CM for Central Europe, and EC-EARTH for the Eastern Mediterranean. At the opposite end of the spectrum, a subset of ten GCMs perform below average in all cases. Finally, a few models have good to excellent results in one domain but fail in others: for example, GFDL-ESM2G—the best model over the British Isles and above average over Central Europe—is one of the worst over the Eastern Mediterranean with the median error of over 50%; on the other hand, EC-EARTH, the best model over the Eastern Mediterranean, falls into the inferior half of models over both Central Europe and the British Isles.

Our results can be compared with two studies that validated ensembles of CMIP3/CMIP5 GCMs over the Euro-Atlantic domain. Pastor and Casado (2012) validated 16

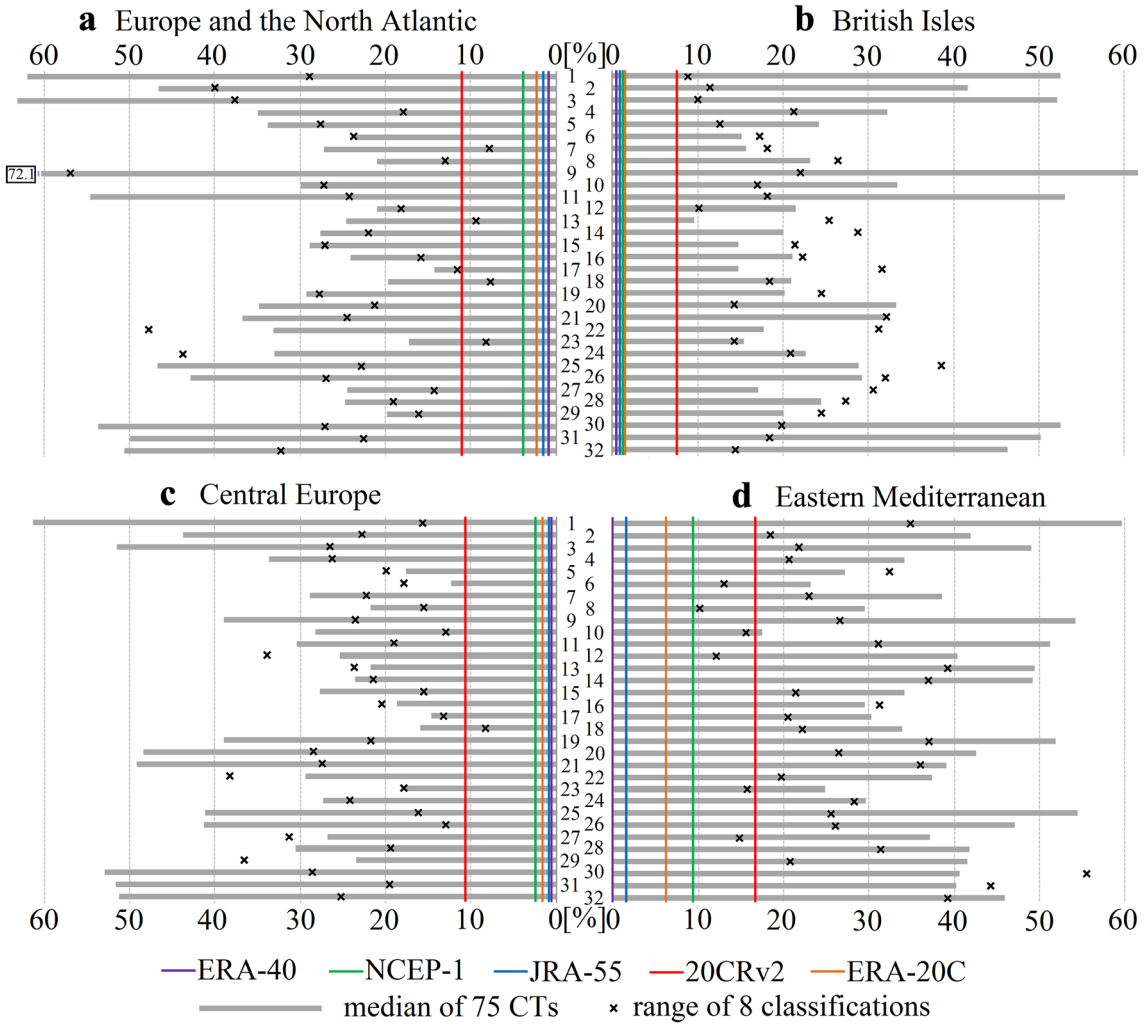


Fig. 2 GCM biases in CT frequency for four European domains. Each horizontal bar refers to one GCM (see Table 2 or Fig. 4 for the GCM legend) and shows the median of absolute values of errors of frequency of 75 CTs. The biases are expressed in percent of the median reanalysis. The crosses indicate the range of median errors (in percent points) if individual classifications are used to compute the median errors instead of all 75 CTs (see Sect. 3.3 for more information and discussion of results). For each reanalysis, the median absolute difference in CT frequency (relative to the median reanalysis) is indicated by a coloured line

CMIP3 GCMs over the Euro-Atlantic region based on two clustering algorithms and winter 1980–1999 data. Most models identified there as best representing ERA-40 CTs frequencies [HadGEM1, GFDL-CM2.0, MIROC3.2(hires) and ECHAM/MPI-OM], were also found among the best here (more precisely, their newer generations). When comparing the results there with the ranking found here for the large domain, sharp differences occur in the rank of some models: CCSM3 was among the best, while CCSM4 is one of the worst here; contrariwise, CNRM-CM3 was among the worst, while CNRM-CM5 is one of the best here. CCSM3 was also found mediocre by Perez et al. (2014) over northeast Atlantic Ocean in all seasons except summer. Naturally, differences can be to some extent expected between different generations of models and for different spatial and temporal domains analysed in different studies. However, we hypothesize that the major cause of the differences is that Pastor and Casado (2012) subtracted SLP biases of the models prior to the classifications. Since CCSM3 had a marked negative SLP bias (− 12 hPa) over central Europe, its subtraction led to more realistic CT frequencies. Without doubt, bias correction can be useful when using GCM simulations. However, models that perform well only if bias corrected should not be considered equal to models that produce accurate climatologies directly, which is exactly what validation of bias-corrected models does.

Perez et al. (2014) validated 42 CMIP5 GCMs over northeast Atlantic Ocean using *k*-means classification of PCs inferred from 1950 to 1999 NCEP-1 3-daily mean SLP anomalies. Their ranking for winter corresponds well with

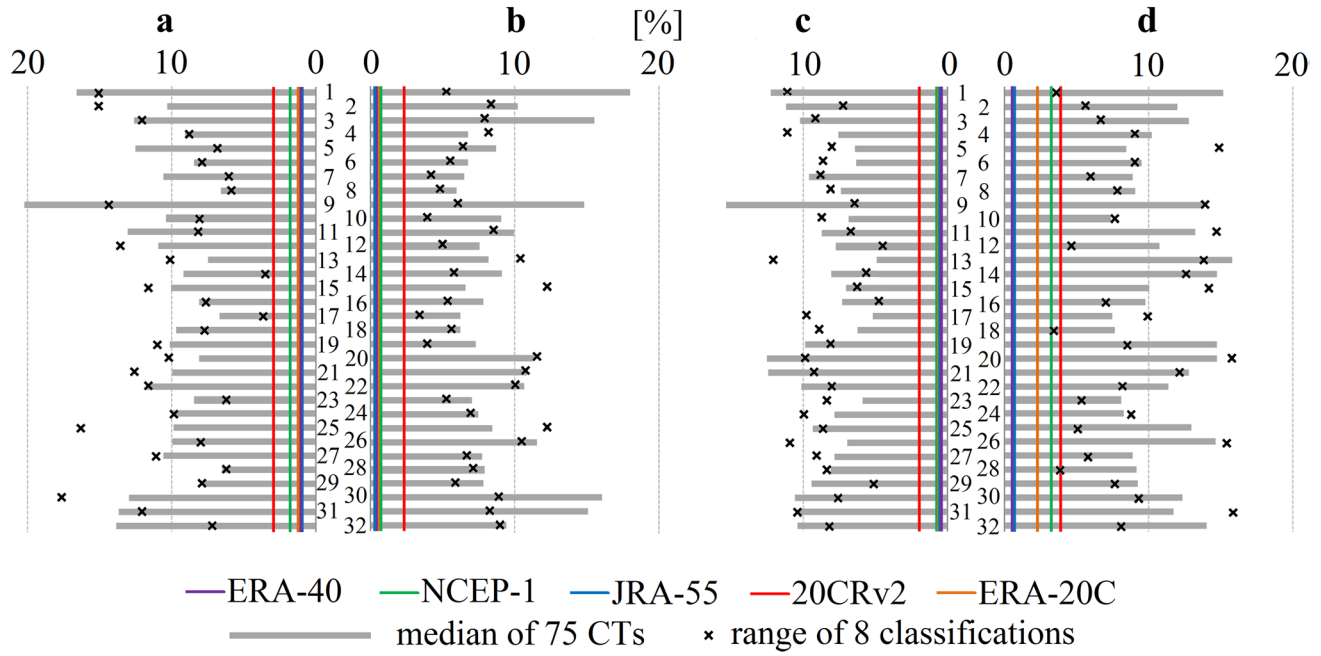


Fig. 3 Same as in Fig. 2, but for persistence. **a** Europe and the North Atlantic, **b** British Isles, **c** Central Europe, **d** Eastern Mediterranean

the ranking for the large domain shown here (r^2 based on linear regression is more than 0.75; note that only 31 GCMs used in both studies are included in this assessment). Expectedly, the ability of GCMs to simulate circulation over the northeast Atlantic Ocean is the primary factor contributing to a correct simulation of CTs over the British Isles and Central Europe (r^2 of 0.56 and 0.66, respectively), while it is less relevant for the Eastern Mediterranean (0.30). The relation is also weak between the rankings for the Eastern Mediterranean and the large Euro-Atlantic domain used here (0.40), which confirms that the selection of models performing well over the Eastern Mediterranean requires dedicated research and cannot rely on results for large Euro-Atlantic domains.

The median errors in frequency and persistence shown, respectively, in Figs. 2 and 3 are highly correlated (Pearson correlation is between 0.77 and 0.84 depending on the domain); consequently, the rankings based on them are similar. Nonetheless, they are not the same and pinpointing the best GCMs always leads to somewhat different results. This can be considered a warning against choosing GCMs for further research based on GCM rankings (instead of on their errors) since the ranking can be markedly altered by only minor changes in the underlying methodology. For instance, GFDL-ESM2G ranks 16th in simulating persistence over the British Isles; however, its median error is only about two percent points higher than that of the best model.

It has to be stressed that even in the best models there are CTs with substantial errors in frequency. Figure 5 illustrates this by showing errors in frequency of all 75 CTs in HadGEM2-CC over the large domain. Therefore, studies that focus on synoptic climatology of certain features (e.g. extremes) should not overly rely on overall rankings but rather on the ability of models to simulate CTs that condition the studied features.

The sensitivity of model biases in CT frequency and persistence to resolution of models seems negligible (not shown). The cross-model correlation between the horizontal resolution (total number of grid points) and median errors in frequency and persistence is significant (at $\alpha = 0.05$) only in the case of CT frequency for the Eastern Mediterranean ($r = -0.42$); otherwise, resolution of models explains less than 10% of the inter-model variability. This result is in agreement with findings by Anstey et al. (2013) and Perez et al. (2014) that the link between model resolution and CT errors is weak and regionally variable. We conjecture that the somewhat stronger link over the Eastern Mediterranean is caused by relatively low SLP gradients there, which leads to higher impacts of smaller-scale features on classifications, which are better resolved by models with finer resolution.

3.3 Sensitivity of validation to the choice of a classification method

So far, all presented results were based on the combination of all eight classifications (75 CTs). In this section we analyse whether the results would differ if the classifications were used separately.

First, Fig. 6 shows median absolute errors in CT frequency of the whole GCM ensemble. Based on all CTs, the ensemble medians are—for the four domains and in

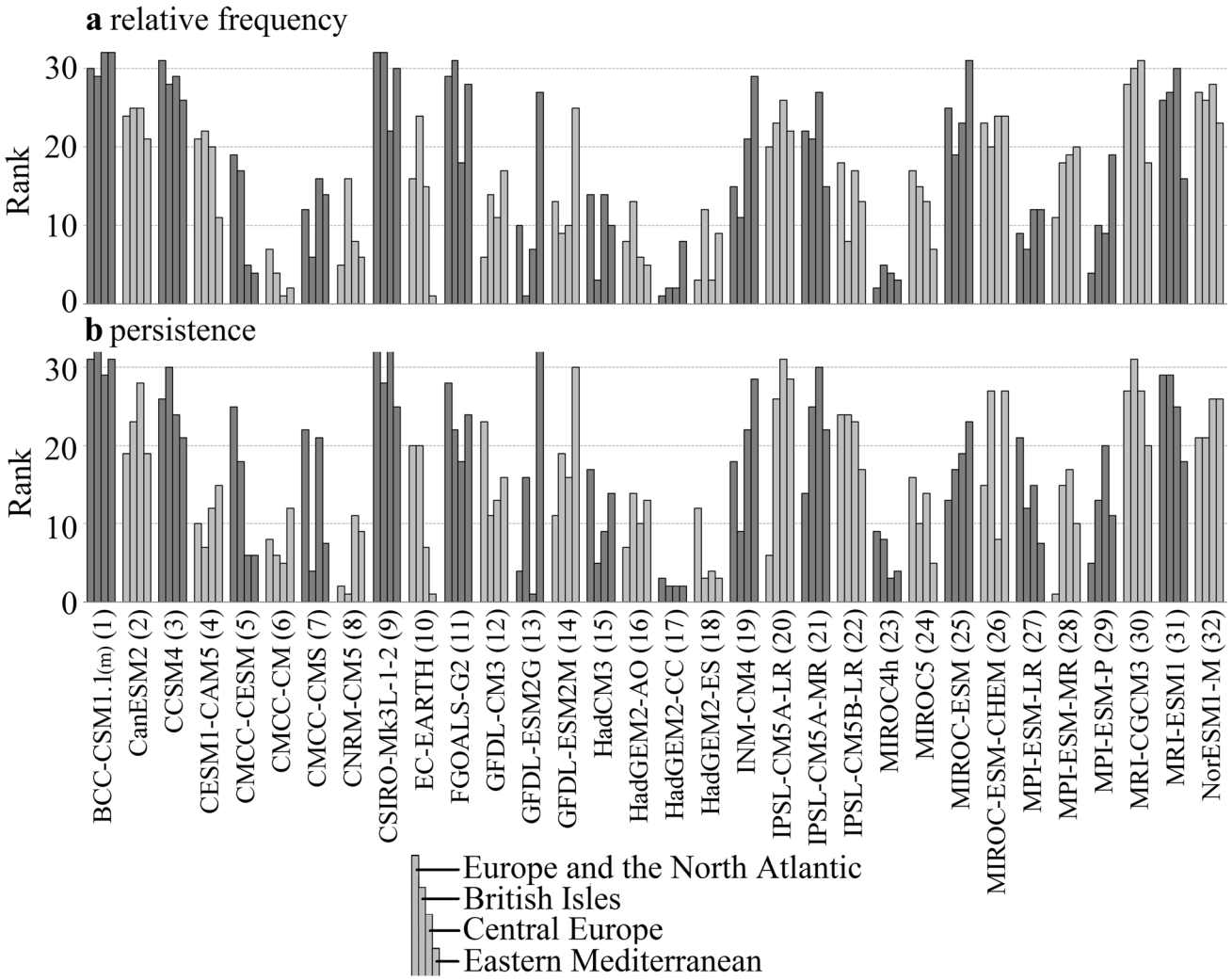


Fig. 4 Rankings of GCMs based on **a** frequency and **b** persistence of CTs for four European domains. The numeric codes in parentheses are the same as in Figs. 2 and 3

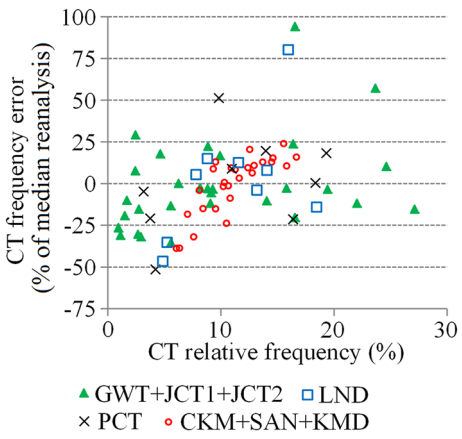


Fig. 5 Errors in frequency of 75 CTs in HadGEM-CC relative to the median reanalysis over the Euro-Atlantic domain. Methodologically similar classifications are grouped together to increase clarity

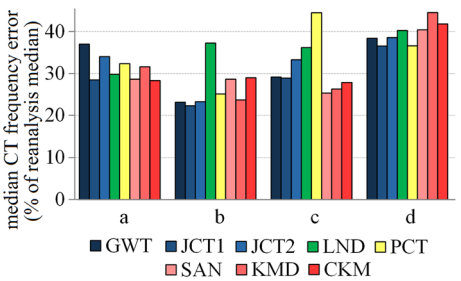


Fig. 6 GCM ensemble biases in CT frequency based on eight classifications for **a** Euro-Atlantic domain, **b** British Isles, **c** Central Europe, and **d** Eastern Mediterranean

the order in which they are shown in Fig. 6—about 33, 24, 30, and 40 percent. Consequently, one would conclude that the skill of the GCM ensemble is best for the British

Isles and worst for the Eastern Mediterranean. However, using only the LND or PCT classification would lead to considerably different conclusions: choosing LND (PCT) instead of JCT1 for the British Isles (Central Europe) would increase the median error by more than 15 percent points, or 60%. Evidently, the result of a comparison of the skill of models over various domains depends on the classification method.

Second, differences between methods become even more apparent if the biases are compared for individual models instead of the ensemble median. The results for the large domain (Fig. 7) indicate that the biases can differ by several hundreds of percent between some classifications: in the extreme case (IPSL-CM5B-LR, #22), they vary between the minimum of about 11.5% for JCT2 and the maximum of 59% for GWT, the range being about 48 percent points. Note that this range was shown for each model and domain in Fig. 2 (for frequency) and Fig. 3 (for persistence) to put it into scale with the overall bias based on all CTs. Expectedly, it is not only the bias of a model what depends on the choice of the classification but also the model's rank (not shown): for example, HadGEM2-CC (#17) ranks 1st based on JCT1 and 6th based on GWT, MPI-ESM-P (#29) ranks between 1st and 13th, and IPSL-CM5B-LR (#22) even between 1st and 27th. Without doubt, basing GCM rankings on one classification is very feeble and should be avoided or at least appropriately interpreted. Namely, such a ranking is not a robust representation of the skill of models to simulate the circulation as a whole but only to simulate a few specific CTs. To conclude, although some effect of the selection of a method on results was expected, it is striking how substantial this effect can be and how much even relatively small changes in the classification procedure (compare e.g. JCT1 and JCT2) alter the results.

3.4 Errors in CT frequency in relation to SLP bias

An alternative—and relatively very easy—way to validate circulation in a GCM output is to compare simulated and reanalysed mean patterns (e.g. compute the bias of the simulated mean SLP pattern). Recently, Wójcik (2015) has shown that SLP biases of GCMs introduce considerable biases into downscaled temperatures and that the overestimated temperature over Europe in winter is caused by GCMs exaggerating the meridional pressure gradient. However, the author concludes that multiyear mean GCM (SLP) biases explain only part of inter-model variability of temperature biases, calling for more in-depth evaluations.

Compared with mean SLP biases, circulation classifications can definitely provide much deeper insight into errors in the simulated circulation. Nevertheless, it seems useful to analyse how much of the inter-model variability of the biases in CT frequencies can be explained by mean SLP biases and whether the classifications are sensitive to the mean SLP bias or rather reflect other, more subtle errors in circulation.

The relation between the mean SLP bias (the bias of a model is defined as the mean difference across all grid points between the two centroids, one for the model and one for all reanalyses) and the bias in the frequency of 75 CTs is for each model and the three small domains illustrated in Fig. 8a. Pearson correlation between the biases in frequency and absolute values of the mean SLP bias (absolute values are used since the sign of the SLP bias is not expected to affect the error in frequency) is relatively low and ranges between 0.23 for the large domain and 0.54 for Central Europe (Table 4, bottom row). The same analysis, except for individual classifications, reveals that the strength of the link considerably varies between the methods. In particular, over the Eastern Mediterranean, classifications by CA (SAN, KMD, CKM) strongly react to the overestimation of SLP (Fig. 8b; Table 4) in contrast to the remaining methods: the coefficients for CA are not only very high ($r_p > 0.85$) and

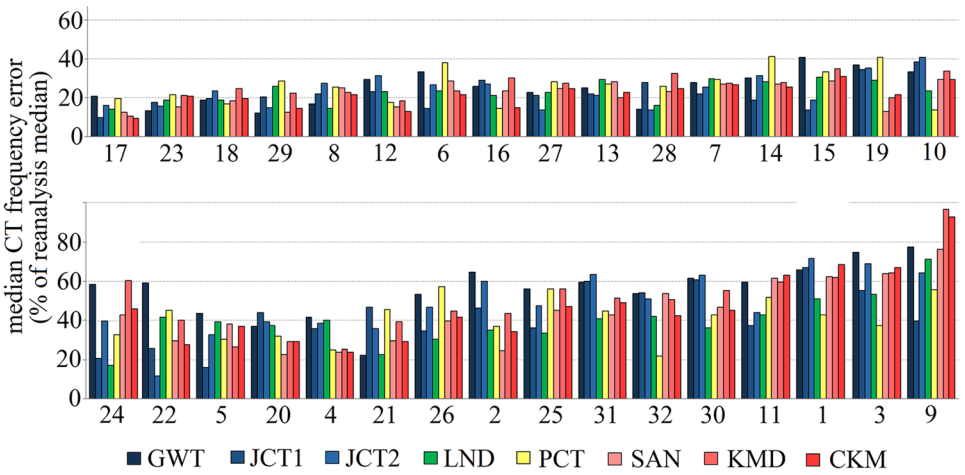


Fig. 7 GCM biases in CT frequency based on eight classifications for the Euro-Atlantic domain. The GCM numeric codes are explained in Table 2. The GCMs are sorted by their overall bias in CT frequency shown in Fig. 2a

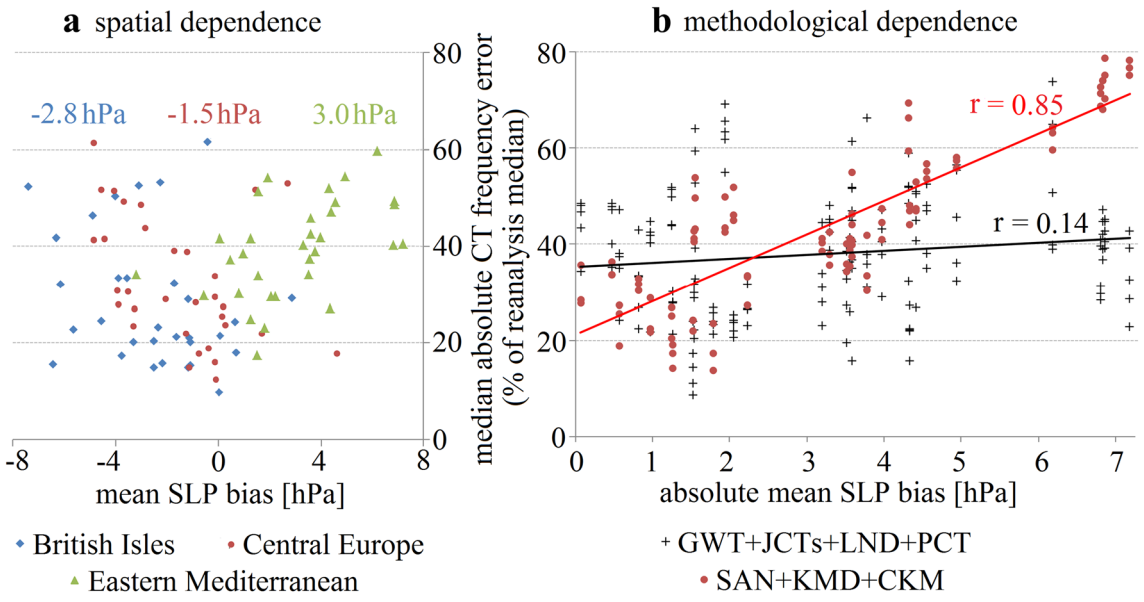


Fig. 8 The relation between GCM biases in mean SLP and CT frequency and its spatial and methodological variation. **a** For every GCM and three domains, the overall bias in CT frequency (i.e. the median error of all 75 CTs) is plotted against the GCM's mean SLP bias. GCM-ensemble mean SLP biases are shown for each domain by numeric values. **b** Same as in **a**, but only for the Eastern Mediterranean, for absolute values of the SLP bias (horizontal axis), and for the CT frequency biases computed and shown for individual classifications instead of all 75 CTs (i.e. eight values are shown for each model). The methods are grouped according to their behaviour: CA-based classifications (SAM, KMD, CKM; red dots) versus the rest (black crosses). Regression line is fitted and and Pearson correlation calculated separately for both groups

Table 4 Pearson correlation coefficients between absolute values of the mean SLP bias and median absolute errors of CT frequency

Method/Domain	Euro-Atlantic	British Isles	Central Europe	Eastern Mediterranean
GWT	0.31*	0.41**	0.32*	−0.07
JCT1	−0.07	0.43**	0.43**	0.27*
JCT2	0.07	0.42**	0.41**	0.12
LND	0.41**	0.36*	0.30*	0.30*
PCT	0.29*	0.36*	0.48**	0.03
SAN	0.38*	0.45**	0.61**	0.85**
KMD	0.22	0.40*	0.60**	0.85**
CKM	0.34*	0.44**	0.66**	0.87**
75 CTs	0.23	0.41**	0.54**	0.49**

*, **The statistical significance of coefficients at the 10% (1%) level (based on one-tailed t-test)

significant but also significantly higher than the coefficients for the remaining methods ($p < 0.001$) [the conventional test for the equality of correlation coefficients is employed (Huth et al. 2006)]. A similar difference between the methods can be seen over Central Europe although the coefficients are more similar for CA and other methods there. For the British Isles and the large domain, the relation between the two biases seems independent of the method.

Over the Eastern Mediterranean, biases in CT frequency in different classifications undoubtedly reflect different kinds of errors in the simulated circulation and, therefore, require a different interpretation. Figure 9 shows selected CTs by SAN to illustrate the behaviour of the method over the domain. When projecting reanalysis centroids onto GCMs, the simulated daily patterns tend to be assigned to the CTs with the highest SLP. In the SAN classification, the circulation type #2 (CT02) is overestimated by nearly 300% (see the left part of Fig. 9). Conversely, if CTs are defined in GCMs and projected onto reanalyses, reanalysed daily patterns tend to be assigned to CTs with overall lowest SLP (right part of Fig. 9). Relying only on this classification and one direction of projection, one could easily arrive at a wrong conclusion

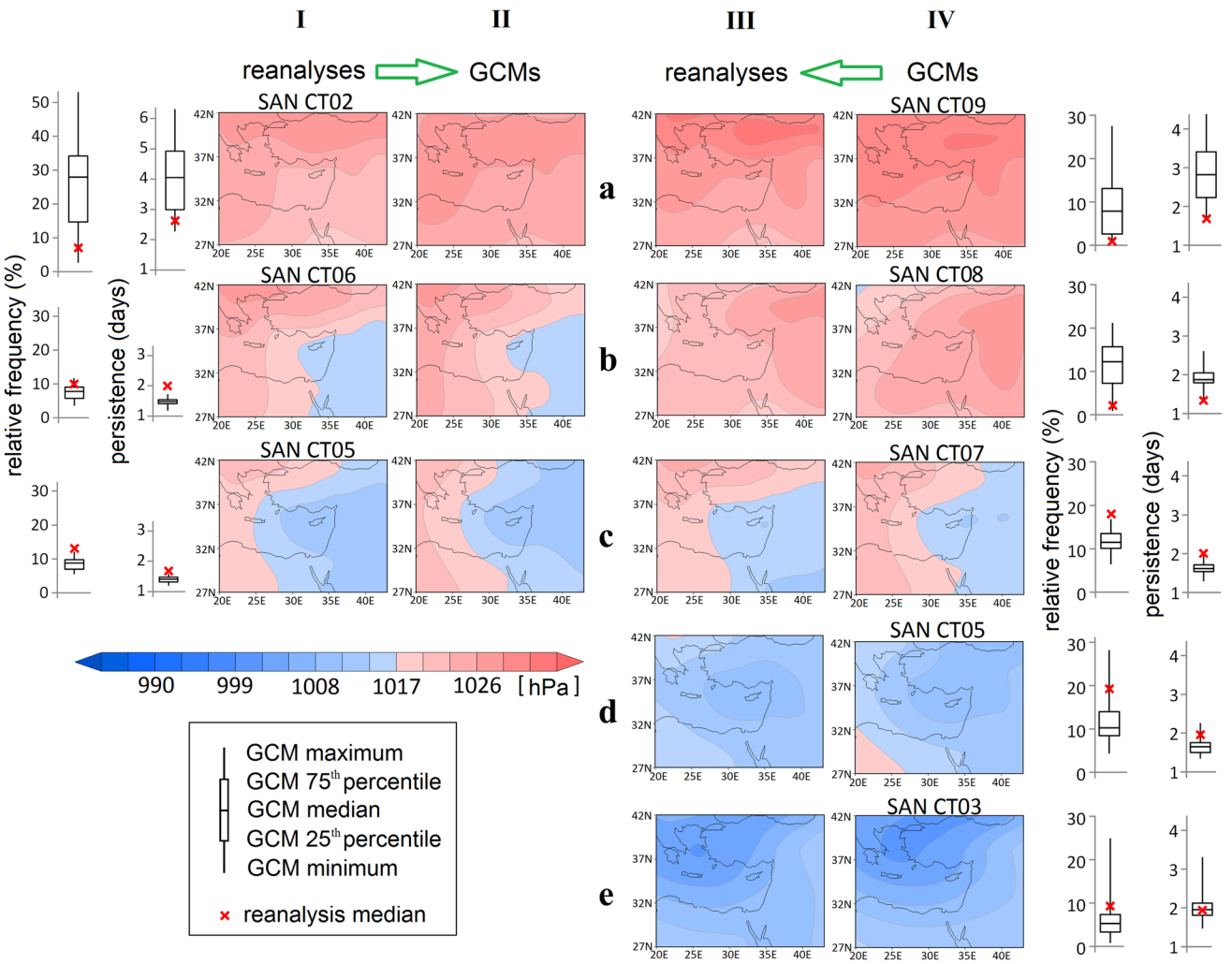


Fig. 9 Two classifications by SAN for the Eastern Mediterranean differing in the direction of projection. In rows **a–e**, selected CTs are shown. Columns **I** and **II** and the pair of boxplots on the left illustrate the projection from reanalyses onto GCMs, columns **III** and **IV** and the pair of boxplots on the right illustrate the projection from GCMs onto reanalyses. The results for the the projection from reanalyses onto GCMs read as follows: The centroids (shown in **I**) are computed from all reanalysed patterns classified with the particular CTs. These centroids are projected onto GCMs and the centroids computed from all simulated patterns that are assigned to the particular CT are shown in **II**. The pairs of leftmost and rightmost graphs show the relative frequency (in percent, left in the pair) and persistence (in days, right in the pair) of the CT in the GCM ensemble (boxplot) and in the median reanalysis (red cross)

that the GCMs severely overestimate the frequency and persistence of advection from E and NE, which—as will be shown later—is directly opposite to the findings based on the remaining methods.

We hypothesize that the reason why the classifications based on cluster analysis strongly respond to SLP bias over the Eastern Mediterranean—and not over the British Isles where the mean SLP bias of the GCM ensemble is similar except for the opposite sign—is that there are markedly weaker SLP horizontal gradients over the Eastern Mediterranean. Therefore, simulated patterns with high mean SLP are often closer (in terms of Euclidean distance) to the centroid with the highest mean SLP regardless of its shape. Consequently, since the classifications do not account for the shape of patterns, errors in CT frequency cannot be interpreted as errors in the direction of flow.

The predominantly weak relation between the biases in mean SLP and in CT frequency may seem to contradict Demuzere et al. (2009) who found that subtracting monthly mean SLP biases of ECHAM5-MPI/OM from ERA-40 and classifying the bias-corrected patterns instead of the original data leads to markedly more realistic CT frequencies. The contradiction is clearly caused by differences in methodology. Foremost, Demuzere et al. (2009) subtracted the mean monthly bias, which is a pattern, and, thus, accounted for the spatial structure of the bias. Here, the spatially aggregated absolute error disregards the spatial structure; consequently, the results illustrate only whether (and how much) the errors

in CT frequency relate to the overall overestimation or underestimation of SLP. Hypothetically, computing correlations between biases in SLP and CT frequencies might be an efficient and simple enough method to test how differences in CT frequencies between datasets should (not) be interpreted. However, these results should be seen as preliminary and requiring verification. For now, to avoid erroneous interpretations, one has to resort to arduous—and therefore usually neglected—steps such as those adopted in this paper: comparing results of multiple classification methods (which itself might not be sufficient should these methods be too similar) and comparing projections in opposite directions.

3.5 GCM-ensemble biases in regional circulation

In this section, the ability of the GCM ensemble median to simulate properties of individual CTs and of main types of circulation (defined by the quadrant of airflow and vorticity) over the four domains is analysed. The results are organised as follows:

First, two indexes are calculated for the centroid of each CT, namely the direction of flow and vorticity, using the algorithm identical to that of the JCT1 classification [refer to Jones et al. (2013) for the formulae of the indexes]. The indexes are calculated separately for the output of all reanalyses and the output of all GCMs.

Second, these indexes are used to "meta-classify" CTs into five to six groups: three to four for directional CTs (approximately based on the quadrant of flow), one for cyclonic (C), and one for anticyclonic (AC) CTs. Note that the boundaries of these groups are not purely objective but were slightly modified to group similar (both in terms of their centroids and if possible also errors in circulation) CTs together. Consequently, they slightly vary between the domains. Based on a visual assessment of CT centroids, the border vorticity of (anti)cyclonic CTs was set to (−)11 hPa for the small domains and (−)20 hPa for the large domain. Figure 10 indicates the boundaries of the groups and shows the GCM biases in CT frequency as a function of the two indexes.

Third, the following statistics are computed for each group and domain: the number of CTs pertaining to a group (regardless of the classification), the range of relative frequencies that a group has in individual classifications, the range of errors in CT frequency across all CTs pertaining to a group, and the median of these errors (bias of CT frequency). In the same way, also the bias of persistence is computed. All these results are shown in Table 5.

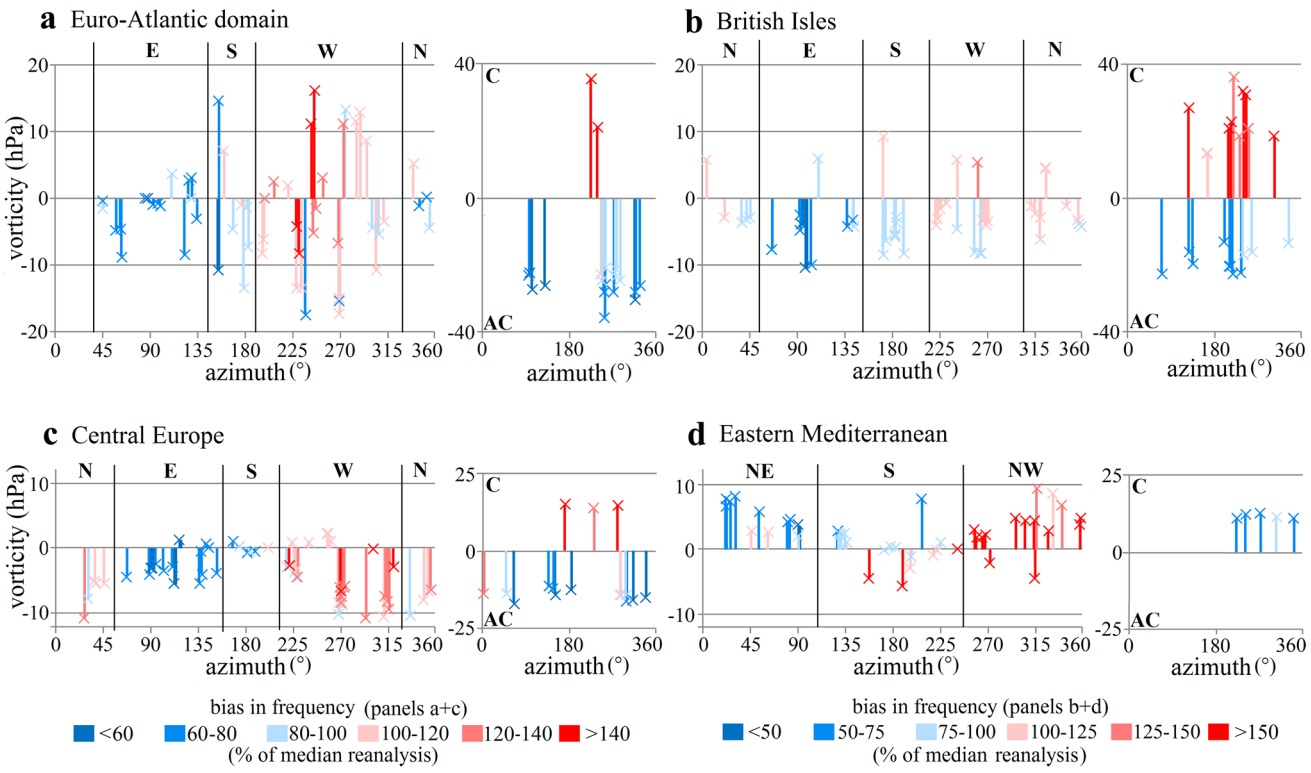


Fig. 10 Biases in CT frequency (GCM median minus reanalysis median; shown in colour: underestimation in blue, overestimation in red) as a function of vorticity (vertical axis) and direction of flow (horizontal axis). All 75 CTs are plotted for the domains shown in **a**–**c**. In **d**, CTs by clustering methods (SAN + CKM + KMD) are not included. Note that the scale of the vertical axes is not uniform

Table 5 Groups of CTs: frequency and biases in simulated frequency and persistence

Euro-Atlantic domain						British Isles					
group	No of CTs	freq[1]	freq errors[2]	freq bias[3]	pers bias[4]	group	No of CTs	freq[1]	freq errors[2]	freq bias[2]	pers bias[4]
W	28	38–71%	-38/+99%	**13%**	-0%	W	15	27–53%	-17/+26%	**7%**	**-3%**
N	4	0–5%	-32/+8%	-21%	**-1%**	N	15	8–28%	-15/+23%	6%	0%
E	16	0–18%	-51/-12%	**-29%**	1%	E	11	5–9%	-52/-24%	**-36%**	**-14%**
S	8	0–25%	-41/+5%	-12%	-4%	S	9	9–23%	-14/+10%	**-9%**	**-9%**
C	2	0–14%	+48/+70%	**59%**	**24%**	C	12	4–21%	+23/+126%	**45%**	3%
AC	17	4–31%	-59/+2%	**-31%**	**-6%**	AC	13	2–18%	-47/-13%	**-36%**	**-7%**
Central Europe						Eastern Mediterranean					
group	No of CTs	freq[1]	freq errors[2]	freq bias[2]	pers bias[4]	group	No of CTs	freq[1]	freq errors[2]	freq bias[2]	pers bias[4]
W	27	48–65%	-7/+107%	**21%**	**4%**	NW	15	13–45%	+23/+271%	**69%**	**14%**
N	9	0–14%	-19/+32%	**5%**	**7%**						
E	15	7–19%	-48/-26%	**-38%**	**-8%**	NE	11	13–18%	-57/+8%	**-34%**	-0%
S	6	0–21%	-33/+6%	-18%	**-11%**	S	17	34–56%	-40/+85%	-7%	**7%**
C	3	0–5%	+34/+46%	**44%**	2%	C	5	4–15%	-49/-18%	**-39%**	-8%
AC	14	4–22%	-63/+27%	**-33%**	**-3%**						

Groups with a non-random (at $\alpha = 0.1$) number (x) of CTs having either underestimated (bold) or overestimated (italics) frequency/persistence in GCMs. The test is as follows: The number of CTs that have an error of the same (and prevailing) sign is a random variable X, which follows a binomial distribution with parameters n equal to the number of CTs in the group and $p = 0.5$ (under H_0, positive and negative errors are assumed to be symmetrically distributed around zero). H_0 is rejected when $P(X \geq x)$ is less than the test level

[1]The range of the relative frequency of days classified with a group in individual classifications

[2]The range of errors of CT frequency across all CTs pertaining to a group

[3,4]The bias is defined as the median of errors of frequency (persistence) of all CTs pertaining to a group

Last, separately for each domain, Fig. 11, 12, 13 and 14 show reanalysis centroids for all CTs, biases of the GCM ensemble in frequency, persistence, and horizontal SLP gradient, and two measures quantifying the correspondence between the reanalysis and GCM centroids (Pearson spatial correlation coefficient and Euclidean distance). The centroids are organised according to their mutual similarity in a two-dimensional array (a "meta-map"). The organisation is subjective, nevertheless, guided by the direction of flow, vorticity, and mutual spatial correlation. Note that since the biases in GCMs found in the classifications by CA for the Eastern Mediterranean cannot be interpreted as errors in the direction of flow (see Sect. 3.4), the three CA-based classifications are excluded from the assessment for this domain.

The range of CT errors in frequency (Table 5) and persistence (not shown) is quite wide for most groups, and the sign of these errors is rarely uniform. Over all domains, westerly CTs tend to occur more often in GCMs (indicated by red colour in the "freq bias" column of Table 5) and the bias of the Wq group increases southeastwards from 7% over the British Isles to almost 70% over the Eastern Mediterranean. Nevertheless, not all of these CTs are overestimated. First, over the British Isles, westerly anticyclonic types obtained by CA methods are underestimated whereas westerly cyclonic types obtained by CA methods are underestimated (as an example, compare two CTs by SAN: anticyclonic #66 versus cyclonic #50; Fig. 12). Second, over Central Europe, CTs in classifications by CA with approximately southwesterly advection are slightly underestimated contrary to other quasi-zonal patterns, the frequency of which is overestimated (Fig. 13; compare CTs by SAN: #51 versus #34, #42, and #48). Third, over the large domain, some CTs with a ridge extending from SW and lows centred over or north of Scandinavia are underestimated (Fig. 11; e.g. #54 by GWT); moreover, the underestimation of westerly CTs with the polar front farther to the north (bottom row in Fig. 11; grouped with AC CTs) corresponds well with the underestimation of similar CTs over the British Isles. On the contrary, the underestimation of some northwesterly CTs over the large domain (Fig. 11; e.g. #46, #54, and #61) is not apparent over Central Europe where northwesterly directional CTs are among the most overestimated in GCMs (Fig. 13; e.g. #31–33). This is a clear warning not to infer errors in regional-scale circulation from classifications—and especially single CTs—constructed for very large domains. Last, the most frequent CT by LND is slightly underestimated over all domains. Since the method tends to classify all patterns highly correlating with the climatological mean with one CT (see the boxplots: #73 in Fig. 11, #51 in Fig. 12, #73 in Fig. 13, and #40 in Fig. 14), the underestimation of this CT in GCMs could mean that the simulations tend to yield patterns more distant from the mean. However, it might also be an artifact of the LND method as similar (in terms of shape and frequency) CTs by GWT and PCT have positive biases of frequency.

Furthermore, groups of easterly CTs (see Fig. 10; Table 5, and patterns with black frames in Figs. 11, 12, 13, 14) have negatively biased frequency over all domains (by about 30–40%) and persistence over the British Isles (14%) and Central Europe (8%). Unlike the westerly CTs, which have a wide range of errors, all easterly CTs are strongly

underestimated, except for a few CTs over the Eastern Mediterranean (see the well-simulated #2 and #4 by JCT1 and #5 by LND in Fig. 14).

The agreement in the errors in frequency of meridional CTs is less clear. Northerly (southerly) CTs usually occur more (less) often over the British Isles and Central Europe. Over the large domain, both groups are slightly underestimated, and the errors in southerly CTs suggest that in GCMs, maritime air masses are more frequently advected into Europe from S to SW at the expense of advection of continental air masses from S to SE (Fig. 10a). This is also apparent over Central Europe where GCMs simulate somewhat stronger flow (#17, #44, #52, and #53 in Fig. 13a, e, f) and lower SLP in southerly CTs, which can also be linked to a vast overestimation of cyclonic CTs over the British Isles ("C" in Fig. 10b).

Not only cyclonic CTs but all CTs with strong vorticity of either sign are poorly captured by GCMs. Anticyclonic CTs are underestimated by more than 30% over all domains except the Eastern Mediterranean where no anticyclonic CTs occur. The only anticyclonic CTs with a correctly simulated frequency are CTs by CA classifications in Central Europe with an anticyclone over or near the Alps (Fig. 13: #54, #61, and #68), although they exaggerate the meridional SLP gradient. Conversely, cyclonic CTs are about half more frequent in simulations except for the Eastern Mediterranean where they have an opposite bias of a similar size. This exception is likely caused by the overestimated SLP in GCMs over the whole Mediterranean region and the considerably more frequent, persistent, and also markedly stronger zonal circulation than in reanalyses (median error of the horizontal SLP gradient in CTs in the NW group is about 30%; see Fig. 14e). The underestimation of CTs with cyclones or troughs over the eastern Mediterranean Sea, Red Sea, and the Middle East has a potential to influence not only local climate, as positive vorticity over southeastern Europe is an important factor of cold spells over central Europe (Buehler et al. 2011; Pfahl et al. 2014). Consequently, the marked underestimation of easterly CTs over Central Europe by GCMs could be interpreted as a result of both the erroneous simulation of Mediterranean circulation and the exaggerated zonality of circulation over the North Atlantic.

Previous study by Pastor and Casado (2012) suggested an overestimation of frequency of occurrence of an AC type with high pressure over western and central Europe and a SW directional CT at the expense of a W directional CT. This is not in agreement with the results here and the discrepancy is likely due to the correction of SLP bias that Pastor and Casado (2012) did and, in particular, of the negative SLP bias over central Europe that they said was present in most of the validated CMIP3 GCMs. Contrariwise, the more frequent and somewhat stronger westerly zonal (and cyclonic) circulation over European domains found here corroborates results of other studies that utilised diverse other approaches to validate circulation in GCMs, such as analyses of the position and intensity of storm tracks and cyclones (Zappa et al. 2013) and the jet stream (Cattiaux et al. 2013a), atmospheric blocking (Vial and Osborn 2012; Dunn-Sigouin and Son 2013), mean patterns (van Ulden and van Oldenborgh 2006; Brands et al. 2013; Wójcik 2015) and circulation indices (Plavcová and Kyselý 2012; Davini and Cagnazzo 2014).

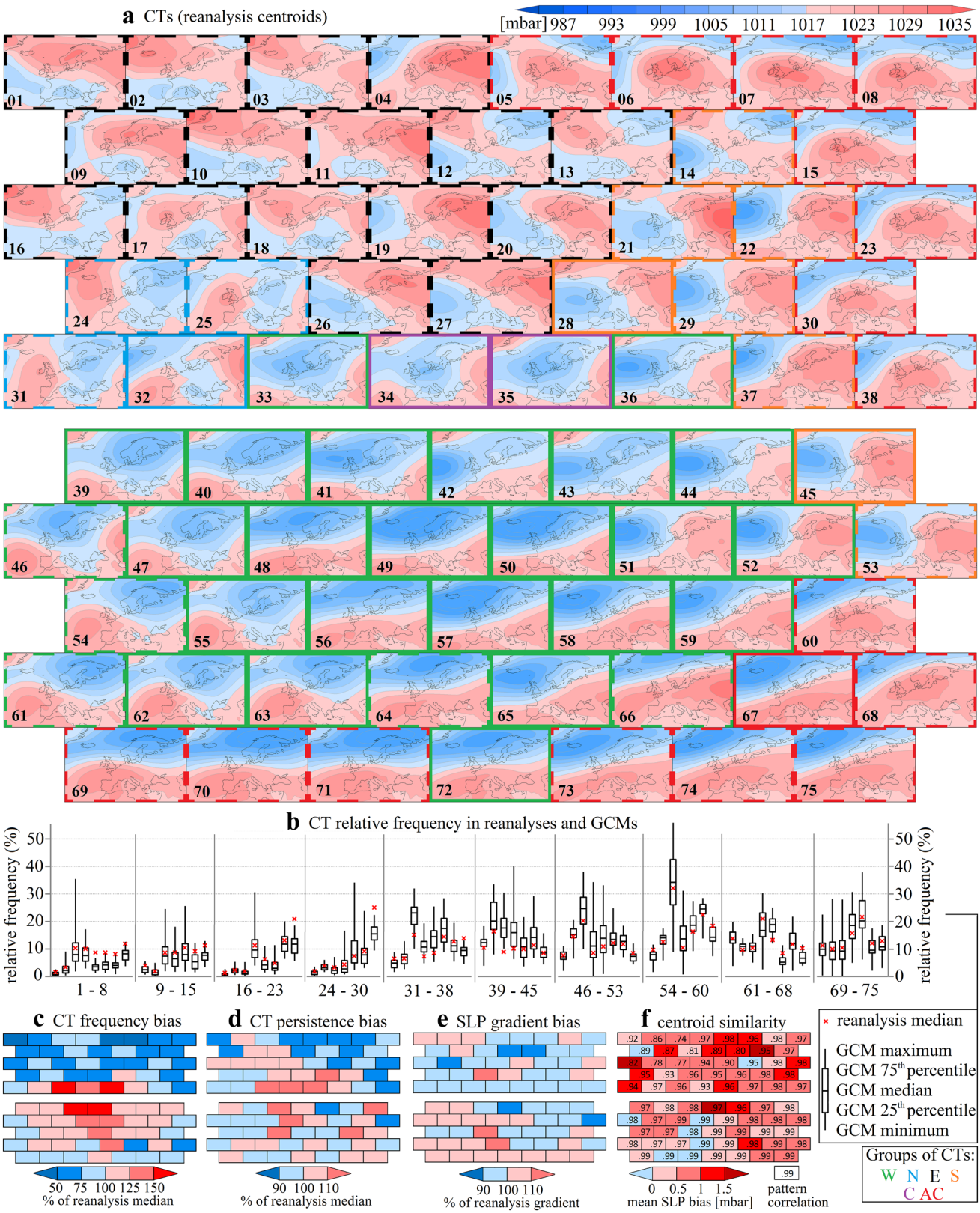
Last, as can be seen in Figs. 11f, 12f, 13f and 14f, CTs for the same domain can considerably differ in their mean SLP bias. Notably, there are differences based on the type of circulation [see e.g. AC CTs in Fig. 13, which overestimate SLP despite their markedly underestimated frequency and although most models underestimate SLP over the domain (Fig. 8a)]. Furthermore, there are also differences between the methods. For example, over the British Isles, CTs by CA methods have almost identical mean SLP in GCMs and reanalyses (the bias is less than 0.7 hPa in absolute value in all 27 CTs), whereas half of CTs by the remaining methods have a negative bias exceeding 2 hPa with the maximum of almost 6 hPa (Fig. 12: #12). On the contrary, correlation coefficients between reanalysis and GCM centroids are generally lower for classifications by CA than the correlations for methods that classify patterns based on correlation (not shown). Therefore, while the latter methods are more suited to investigate how SLP (GPH) differs between CTs in different datasets, the former are better for example for comparisons of SLP gradients.

3.6 Limitations of the study

Despite our effort to obtain as robust a validation result as possible, classifications have some limitations as for their relevance and usefulness; additional limitations stem from subjective choices that were made prior to the study.

Number of circulation types The study is based on a preselected number of circulation types, which is 9–10. This number is enough to include for example all octants of flow and two vorticity types (as in the case of hybrid methods) and at the same time avoid defining types with negligible occurrences, which have generally very large errors in simulations and thus affect the validation to a degree that exceeds their climatological relevance. However, the approach may fail to define CTs with rare occurrences but potentially high impacts on society. Therefore, the results have to be seen as too coarse for applications such as statistical downscaling of weather extremes. Previous studies indicated that the number of CTs is an important factor; its effect on validation was not tested here.

Climatological relevance of CTs and GCM errors The climatological relevance of individual CTs and thus the significance of errors in these types were not tested.



Instead, the results represent a robust estimate of GCM biases in the synoptic-scale circulation as a whole and one has to keep in mind that model rankings built for specific synoptic-climatological case studies may differ from those presented here if additional variables were included and, for example, different weights were assigned to individual

◂**Fig. 11** Characteristics of CTs and their errors in the Euro-Atlantic domain. **a** Centroids of 75 CTs by eight classifications are organised into a "meta-map" based on the similarity of their patterns. The groups of CTs (as shown in Fig. 10; Table 5) are highlighted here by coloured frames with solid (dashed) lines indicating overestimated (underestimated) frequency of CTs in GCMs. **b** Boxplots of CT frequency in GCMs and CT frequency in the median reanalysis (red crosses). **c** GCM-ensemble bias in CT frequency (difference between the median frequency of the GCM ensemble and the median frequency of the reanalysis ensemble). **d** Same as **c**, but for persistence. **e** Same as **c**, but for the horizontal SLP gradient. **f** Correspondence of reanalysis and GCM CT centroids expressed as pattern correlation (numbers) and mean SLP bias (colours). The panels are split in half to improve readability

classifications based on their ability to stratify the given variables. Furthermore, only the most obvious weaknesses of the used methods are discussed, for example those stemming from clustering of raw SLP data.

Selection of methods The sensitivity of validation to the choice of the method implies that our results are to some extent affected by the choice of the ensemble of classification methods. The selection was primarily constrained by the time-consuming nature of producing, analysing, and interpreting classifications. Furthermore, methods of hierarchical CA were excluded since they do not allow for projection (Huth 2000). Self-organizing maps (SOMs) were omitted for two reasons: first, presumably the main benefits of SOMs—visualisation of CTs and their organisation into a rectangular array (see e.g. Sheridan and Lee 2011)—would not have been utilised owing to the amount of other classifications included in the study and the way how results are presented. Second, Philipp et al. (2016) documented that SOMs tend to lead to classifications very similar to SANDRA, a method that was used here.

Projection of CTs The utilization of projection considerably increased the effectivity of the validation as projection circumvents the necessity to define CTs in all datasets combined. However, it also disregards any existing differences in the structure of the two compared datasets. We accounted for this drawback by two-way projections. Nevertheless, in the combined output of all GCMs, it was not possible to compute the classifications by LND due to excessive RAM requirements and KMD due to excessive computation time; therefore, only PCT, SAN, and CKM were verified this way.

The choice of the variable and the type of data The study validated only SLP patterns. Somewhat different results may have been obtained for other circulation variables, for example 500 hPa GPH. Furthermore, only raw SLP data were classified; another option would be to classify SLP anomalies instead. Arguably, classifying anomalies would have weakened or even removed the issue found for clustering of Eastern-Mediterranean patterns that caused the classifications to respond to errors in mean SLP rather than errors in the shape of patterns. Since over-/underestimation of SLP over a larger region can be interpreted as an error in planetary-scale rather than synoptic-scale circulation, clustering of raw data provides a qualitatively different kind of validation than clustering of anomalies or bias-corrected patterns. Consequently, validation of SLP anomalies might be less sensitive to the choice of the method than what we found for raw data. Similarly, validation of GPHs might be less sensitive to the choice of the method as well, because SLP patterns are far more complex and different statistical approaches very likely vary in their response to this complexity. Additional research will be necessary do address the issue of the choices of circulation variable and data type in validation studies.

4 Summary and conclusions

Winter daily SLP patterns over the Euro-Atlantic region in 1961–2000 were analysed in historical runs of 32 CMIP5 GCMs by means of automated circulation classifications. Previous research made clear that the choice of the classification method is an important factor in any synoptic-climatological study. Therefore, eight classifications (each comprising 9–10 CTs, making up 75 CTs in total) were produced in five atmospheric reanalyses (ERA-40, NCEP-1, JRA-55, 20CRv2, and ERA-20C) for the large Euro-Atlantic domain and three smaller domains within its range (British Isles, Central Europe, and Eastern Mediterranean) to minimize the risk of wrong assessments and provide robust estimates of biases in circulation in GCMs. The main results and conclusions are as follows:

The reference dataset was defined as a "median reanalysis", that is, by computing the median of all reanalyses for each CT property [frequency of occurrence, persistence, and mean map (centroid)]. Subsequently, the bias of each model was quantified relative to the median reanalysis as the median absolute error of frequency of all the 75 CTs regardless of the classification. Based on this bias, only HadGEM2-CC, MIROC4h, and CNRM-CM5 appear among the best ten models over all four domains. Over each domain, a different model ranks first (HadGEM2-CC over the large domain, GFDL-ESM2G over the British Isles, CMCC-CM over Central Europe, and EC-EARTH over the Eastern Mediterranean). Rankings for the British Isles and Central Europe relatively highly covariate with the ranking for the large domain and also with a ranking computed by Perez et al. (2014) for the northeastern Atlantic Ocean, while the skill of GCMs over the Eastern Mediterranean is relatively independent of the skill over the other domains. Furthermore, the skill of models considerably varies: the bias of the best model for each domain is nearly the same as the median deviation of the 20CRv2 reanalysis from the median reanalysis, which alone could be considered a very good result;

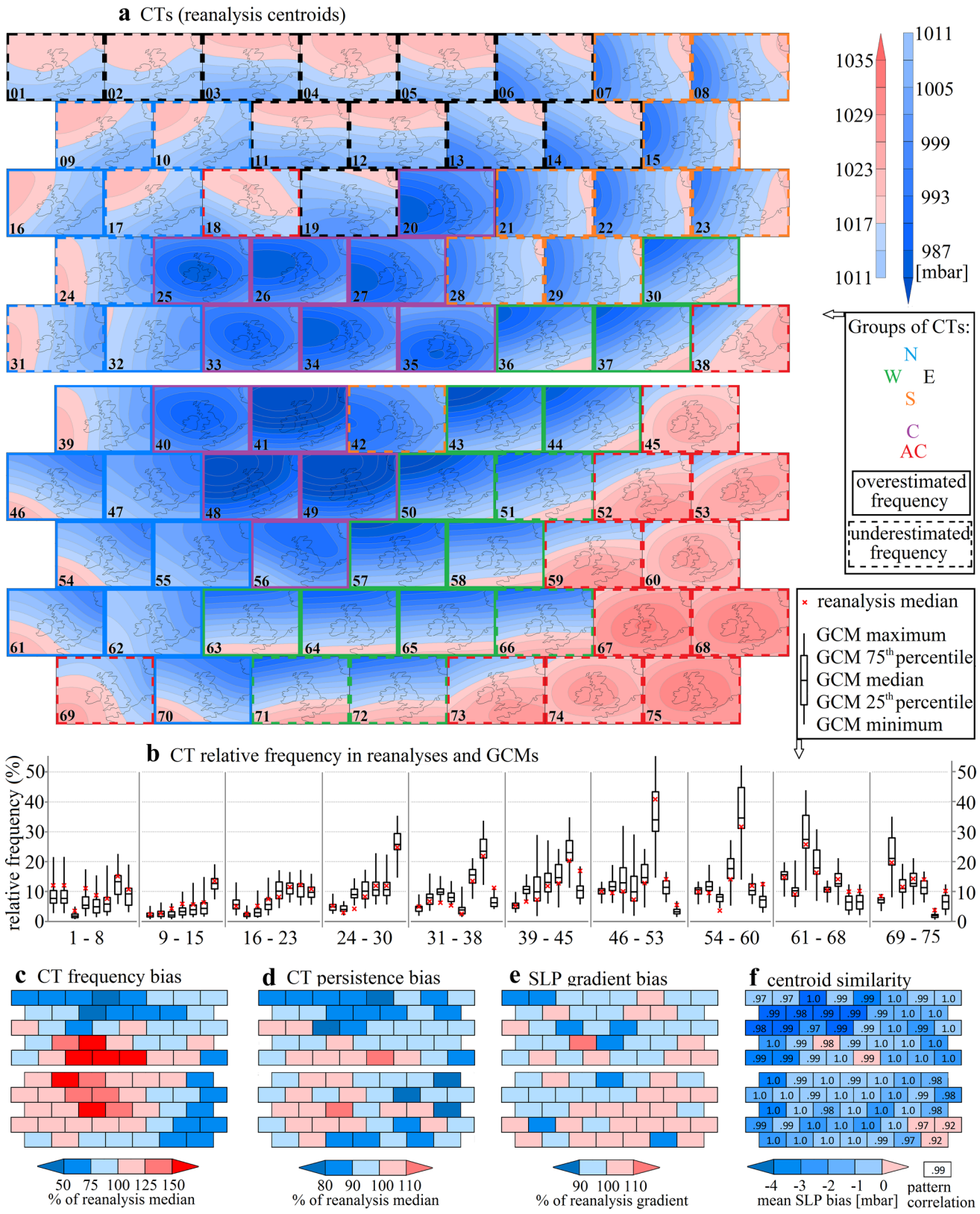


Fig. 12 Same as in Fig. 11, but for the British Isles

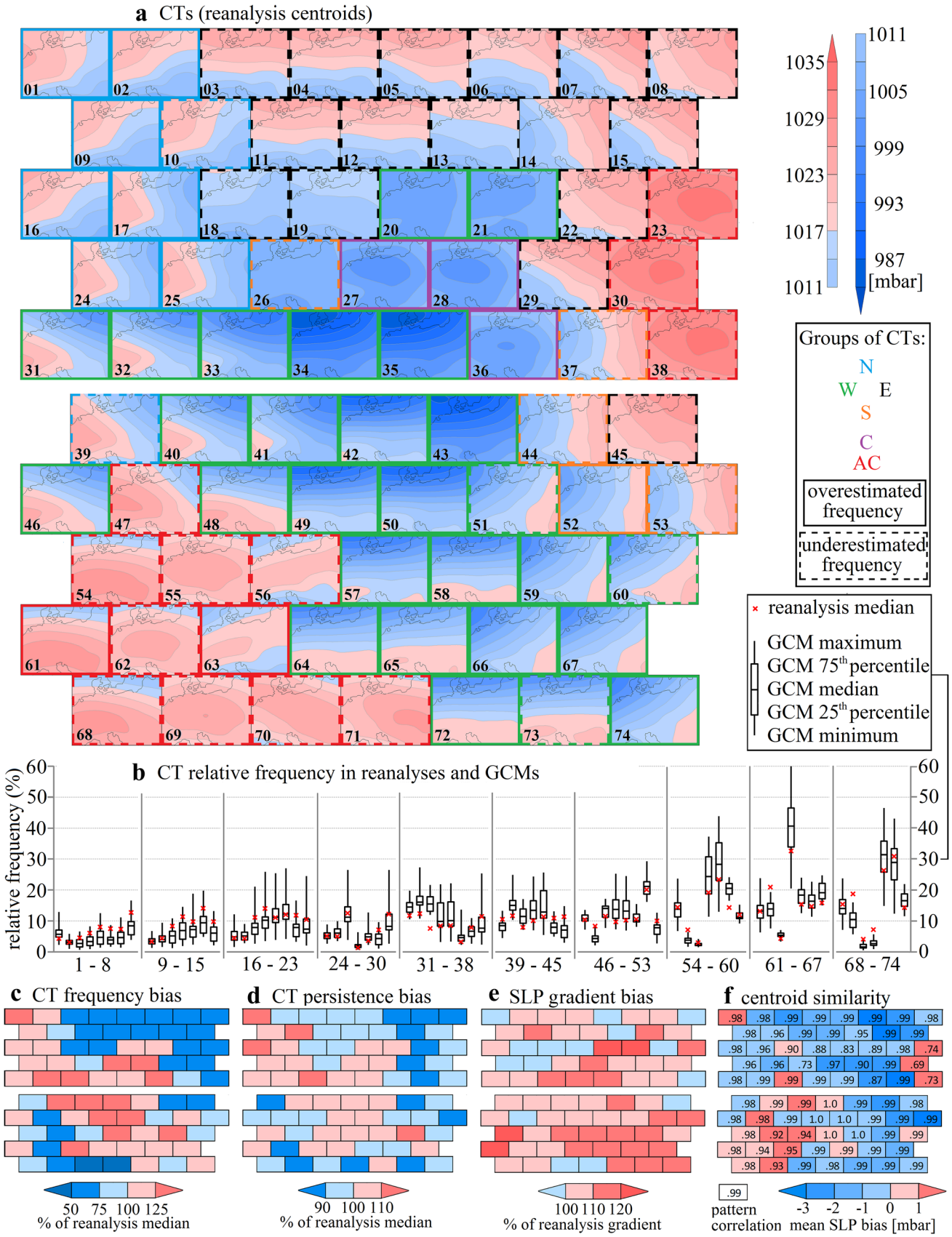


Fig. 13 Same as in Fig. 11, but for Central Europe

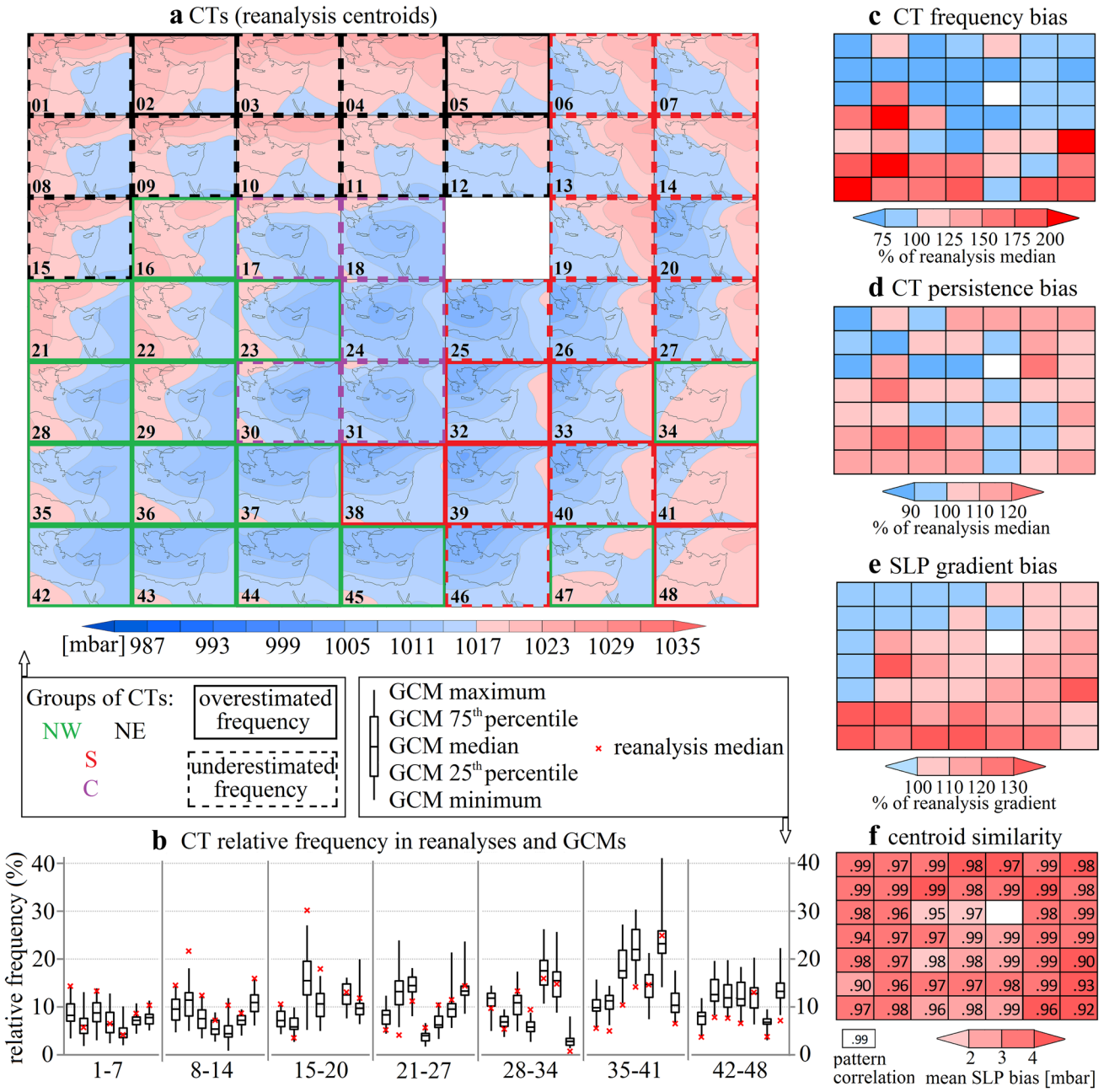


Fig. 14 Same as in Fig. 11, but for the Eastern Mediterranean. Classifications based on cluster analysis (SAN, KMD, CKM) are excluded

however, some CTs can occur as much as twice more often than in reanalyses even in the best models. In contrast, the worst models have a bias in CT frequency of 50% and more. The persistence of CTs is simulated considerably better by the GCMs, most CTs having errors smaller than 15% of the median reanalysis.

Based on all eight classifications combined, the GCM ensemble simulates the CT frequency best over the British Isles (the bias is less than 25%) and worst over the Eastern Mediterranean (40%). However, using the methods separately, one would arrive at contradictory conclusions, which illustrates that the assessment of the skill of models depends on the choice of the method. Moreover, the influence of the method is even stronger on the ranking of models; in the extreme case, IPSL-CM5B-LR ranks first based on one classification and only twenty-seventh (out of 32) based on another. Therefore, it must be remembered that a single classification provides one with only a feeble image of reality and this image must not be overinterpreted. Comparing multiple classifications, on the other hand, helps create

robust estimates of GCM biases, select best GCMs, identify suboptimal classifications, and recognize statistical artifacts.

Demuzere et al. (2009) showed that subtracting mean SLP bias of GCMs leads to considerably more realistic CT frequencies. Here, the link between models' mean SLP biases and median absolute CT frequency errors appears significant in most classifications, except for the Euro-Atlantic domain, arguably due to its size. Over the Eastern Mediterranean, CT frequency biases in classifications by cluster analysis—which use the Euclidean distance as the similarity measure—strongly depend on the mean SLP bias of the models. Consequently, the errors in CT frequencies cannot be interpreted as errors in the direction of flow since it is not similarity in the shape of patterns but rather in the mean SLP what governs the assignment of patterns to CTs. This issue, already investigated by Belleflamme et al. (2013), should theoretically be minimized by classifying patterns of SLP anomalies instead of SLP patterns and/or subtracting SLP bias prior to classification. However, further research seems necessary to verify this and to shed more light on the interpretability of differences between classifications in two datasets if Euclidean distance based methods are used. Here, a simple measure was proposed that consists in computing correlations between biases in CT frequency and mean SLP bias, which should be able to identify the cases when the classification disregards the shape of patterns.

The link between model horizontal resolution and model errors in CT frequency and persistence was investigated. It is mostly weak; it is significant only for the frequency of CTs over the Eastern Mediterranean, where resolution explains about 20% of the inter-model variability of errors. In all other cases, less than 10% of the inter-model variability can be explained by resolution.

Regarding the regional circulation, GCM biases depend on the direction and vorticity of airflow. First, we demonstrate that westerly circulation is overestimated over all domains; the bias ranges from about 7% over the British Isles to almost 70% over the Eastern Mediterranean. The frequency of only a few westerly CTs is underestimated; this concerns CTs with a ridge extending from the Azores across Europe and those with the polar front farther in the north over the large domain, the CTs with marked anticyclonic vorticity over the British Isles, and some CTs with southwest advection over Central Europe. Second, easterly circulation is less frequent in the ensemble than in reanalyses by about 30–40% over all domains and significantly less persistent over the British Isles and Central Europe. Third, northerly circulation tends to be more frequent over the British Isles and Central Europe, while southerly types occur less often in simulations there. Over the large domain, meridional circulation is slightly underestimated and models tend to prioritise zonal advection of maritime air masses onto the continent. Last, circulation with strong vorticity—both cyclonic and anticyclonic—is poorly simulated by the GCM ensemble. Anticyclonic CTs are underestimated by more than 30% over all domains whereas cyclonic CTs are about half more frequent in simulations than in reanalyses. The only exception is the Eastern Mediterranean where anticyclonic CTs do not occur and the frequency of cyclonic CTs is underestimated, probably on account of models considerably overestimating SLP and the meridional SLP gradient over the region.

The present study enhances previous research into validation of circulation in GCM simulations mainly by using multiple methods. One has to keep in mind that in spite of this enhancement model errors (and rankings) are sensitive to various factors (such as number of CTs, selection of methods, reference dataset, spatial and temporal domains, and variables), some of which were not tested here. Furthermore, the validation is carried out for a period of time that is short relative to some natural oscillations (e.g. Atlantic Multidecadal Oscillation). Additionally, only CT frequency and persistence are validated; further research is necessary to address the simulation of the link between CTs and surface weather. Considering all these limitations, neither the model rankings presented here nor the ability of the pinpointed best models to produce reliable projections of future climate should be overrated. The results may nevertheless be used to weigh projections by multi-model ensembles or guide the selection of a subset of GCMs for dynamical downscaling.

Acknowledgements The work was funded by the Grant Agency of Charles University, Project No. 188214. We thank all climate-modelling groups for making available their GCM simulations and the PCMDI for enabling access to the data. We acknowledge the following organizations for providing their reanalysis datasets: NOAA/OAR/ESRL PSD, Boulder, Colorado for the NCEP/NCAR reanalysis and the Twentieth Century Reanalysis, version 2; ECMWF for ERA-40 and ERA-20C; and JMA for JRA-55. Thanks are also due to all developers of the COST733 software and namely Dr Andreas Philipp from the Institute of Geography, University of Augsburg, Germany for the many instructions on its usage; the Institute is also acknowledged for maintaining the software and enabling access to it. We are grateful to two reviewers, whose insight helped improve the quality of this paper.

References

Anstey JA, Davini P, Gray LJ, Woollings TJ, Butchart N, Cagnazzo C, Christiansen B, Hardiman SC, Osprey SM, Yang S (2013) Multi-model analysis of Northern Hemisphere winter blocking: model biases and the role of resolution. J Geophys Res Atmos 118:3956–3971. https://doi.org/10.1002/jgrd.50231

Beck C, Jacobeit J, Jones PD (2007) Frequency and within-type variations of large-scale circulation types and their effects on low-frequency climate variability in central Europe since 1780. Int J Climatol 27:473–491. https://doi.org/10.1002/joc.1410

Beck C, Weitnauer C, Jacobeit J (2014) Downscaling of monthly PM10 indices at different sites in Bavaria (Germany) based on circulation type classifications. Atmos Pollut Res 5:741–752. https://doi.org/10.5094/APR.2014.083

Beck C, Philipp A, Jacobeit J (2015) Interannual drought index variations in Central Europe related to the large-scale atmospheric circulation—application and evaluation of statistical downscaling approaches based on circulation type classifications. Theor Appl Climatol 121:713–732. https://doi.org/10.1007/s00704-014-1267-z

Beck C, Philipp A, Streicher F (2016) The effect of domain size on the relationship between circulation type classifications and surface climate. Int J Climatol 36:2692–2709. https://doi.org/10.1002/joc.3688

Belleflamme A, Fettweis X, Lang C, Erpicum M (2013) Current and future atmospheric circulation at 500 hPa over Greenland simulated by the CMIP3 and CMIP5 global models. Clim Dyn 41:2061–2080. https://doi.org/10.1007/s00382-012-1538-2

Belleflamme A, Fettweis X, Erpicum M (2015) Do global warming-induced circulation pattern changes affect temperature and precipitation over Europe during summer? Int J Climatol 35:1484–1499. https://doi.org/10.1002/joc.4070

Boer GJ, McFarlane NA, Lazare M (1992) Greenhouse gas–induced climate change simulated with the CCC second-generation general circulation model. J Clim 5:1045–1077. https://doi.org/10.1175/1520-0442(1992)005<1045:GGCCSW>2.0.CO;2

Brands S, Herrera S, Fernández J, Gutiérrez JM (2013) How well do CMIP5 Earth System Models simulate present climate conditions in Europe and Africa? Clim Dyn 41:803–817. https://doi.org/10.1007/s00382-013-1742-8

Broderick C, Fealy R (2015) An analysis of the synoptic and climatological applicability of circulation type classifications for Ireland. Int J Climatol 35:481–505. https://doi.org/10.1002/joc.3996

Buehler T, Raible CC, Stocker TF (2011) The relationship of winter season North Atlantic blocking frequencies to extreme cold or dry spells in the ERA-40. Tellus A 63:212–222. https://doi.org/10.1111/j.1600-0870.2010.00492.x

Cahynová M, Huth R (2016) Atmospheric circulation influence on climatic trends in Europe: an analysis of circulation type classifications from the COST733 catalogue. Int J Climatol 36:2743–2760. https://doi.org/10.1002/joc.4003

Casado MJ, Pastor MA (2016) Circulation types and winter precipitation in Spain. Int J Climatol 36:2727–2742. https://doi.org/10.1002/joc.3860

Casado MJ, Pastor MA, Doblas-Reyes FJ (2010) Links between circulation types and precipitation over Spain. Phys Chem Earth 35:437–447. https://doi.org/10.1016/j.pce.2009.12.007

Cassano JJ, Uotila P, Lynch A (2006) Changes in synoptic weather patterns in the polar regions in the twentieth and twenty-first centuries, part 1: Arctic. Int J Climatol 26:1027–1049. https://doi.org/10.1002/joc.1306

Cattiaux J, Douville H, Peings Y (2013a) European temperatures in CMIP5: origins of present-day biases and future uncertainties. Clim Dyn 41:2889–2907. https://doi.org/10.1007/s00382-013-1731-y

Cattiaux J, Douville H, Ribes A, Chauvin F, Plante C (2013b) Towards a better understanding of changes in wintertime cold extremes over Europe: a pilot study with CNRM and IPSL atmospheric models. Clim Dyn 40:2433–2445. https://doi.org/10.1007/s00382-012-1436-7

Compo GP et al (2011) The twentieth century reanalysis project. Q J R Meteorol Soc 137:1–28. https://doi.org/10.1002/qj.776

Crane RG, Barry RG (1988) Comparison of the MSL synoptic pressure patterns of the Arctic as observed and simulated by the GISS general circulation model. Meteorol Atmos Phys 39:169–183

Davini P, Cagnazzo C (2014) On the misinterpretation of the North Atlantic Oscillation in CMIP5 models. Clim Dyn 43:1497–1511. https://doi.org/10.1007/s00382-013-1970-y

Demuzere M, Werner M, Van Lipzig N, Roeckner E (2009) An analysis of present and future ECHAM5 pressure fields using a classification of circulation patterns. Int J Climatol 29:1796–1810. https://doi.org/10.1002/joc.1821

Demuzere M, Kassomenos P, Philipp A (2011) The COST733 circulation type classification software: an example for surface ozone concentrations in Central Europe. Theor Appl Climatol 105:143–166. https://doi.org/10.1007/s00704-010-0378-4

Dunn-Sigouin E, Son SW (2013) Northern Hemisphere blocking frequency and duration in the CMIP5 models. J Geophys Res Atmos 118:1179–1188. https://doi.org/10.1002/jgrd.50143

Enke W, Spekat A (1997) Downscaling climate model outputs into local and regional weather elements by classification and regression. Clim Res 8:195–207

Finnis J, Cassano J, Holland M, Uotila P (2009a) Synoptically forced hydroclimatology of major Arctic watersheds in general circulation models; Part 1: the Mackenzie River Basin. Int J Climatol 29:1226–1243. https://doi.org/10.1002/joc.1753

Finnis J, Cassano J, Holland M, Uotila P (2009b) Synoptically forced hydroclimatology of major Arctic watersheds in general circulation models; Part 2: Eurasian watersheds. Int J Climatol 29:1244–1261. https://doi.org/10.1002/joc.1769

Fleig AK, Tallaksen LM, Hisdal H, Stahl K, Hannah DM (2010) Intercomparison of weather and circulation type classifications for hydrological drought development. Phys Chem Earth 35:507–515. https://doi.org/10.1016/j.pce.2009.11.005

Gibson PB, Uotila P, Perkins-Kirkpatrick SE, Alexander LV, Pitman AJ (2016) Evaluating synoptic systems in the CMIP5 climate models over the Australian region. Clim Dyn 47:2235–2251. https://doi.org/10.1007/s00382-015-2961-y

Hall A (2014) Projecting regional change. Science 346:1461–1462. https://doi.org/10.1126/science.aaa0629

Huth R (1996) Properties of the circulation classification scheme based on the rotated principal component analysis. Meteorol Atmos Phys 59:217–233

Huth R (1997) Continental-scale circulation in the UKHI GCM. J Clim 10:1545–1561. https://doi.org/10.1175/1520-0442(1997)010<1545:CSCITU>2.0.CO;2

Huth R (2000) A circulation classification scheme applicable in GCM studies. Theor Appl Climatol 67:1–18. https://doi.org/10.1007/s007040070012

Huth R (2010) Synoptic-climatological applicability of circulation classifications from the COST733 collection: first results. Phys Chem Earth 35:388–394. https://doi.org/10.1016/j.pce.2009.11.013

Huth R, Pokorná L, Bochníček J, Hejda P (2006) Solar cycle effects on modes of low-frequency circulation variability. J Geophys Res 111:D22107. https://doi.org/10.1029/2005JD006813

Huth R, Beck C, Philipp A, Demuzere M, Ustrnul Z, Cahynová M, Kyselý J, Tveito OE (2008) Classifications of atmospheric circulation patterns. Recent advances and applications. Ann N Y Acad Sci 1146:105–152. https://doi.org/10.1196/annals.1446.019

Huth R, Beck C, Tveito OE (2010) Classifications of atmospheric circulation patterns—theory and applications—preface. Phys Chem Earth 35:307–308. https://doi.org/10.1016/j.pce.2010.06.005

Huth R, Beck C, Kučerová M (2016) Synoptic-climatological evaluation of the classifications of atmospheric circulation patterns over Europe. Int J Climatol 36:2710–2726. https://doi.org/10.1002/joc.4546

James PM (2006) An assessment of European synoptic variability in Hadley Centre Global Environmental models based on an objective classification of weather regimes. Clim Dyn 27:215–231. https://doi.org/10.1007/s00382-006-0133-9

Jones PD, Harpham C, Briffa KR (2013) Lamb weather types derived from reanalysis products. Int J Climatol 33:1129–1139. https://doi.org/10.1002/joc.3498

Kalnay E et al. (1996) The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc 77:437–470. https://doi.org/10.1175/1520-0477(1996)077%3C0437:TNYRP%3E2.0.CO;2

Kassomenos P (2010) Synoptic circulation control on wild fire occurrence. Phys Chem Earth 35:544–552. https://doi.org/10.1016/j.pce.2009.11.008

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley Series in probability and mathematical statistics: applied probability and statistics. Wiley, New York

Kobayashi S et al (2015) The JRA-55 Reanalysis: General specifications and basic characteristics. J Meteorol Soc Japan 93:5–48. https://doi.org/10.2151/jmsj.2015-001

Kröner N, Kotlarski S, Fischer E, Lüthi D, Zubler E, Schär C (2017) Separating climate change signals into thermodynamic, lapse-rate and circulation effects: theory and application to the European summer climate. Clim Dyn 48:3425–3440. https://doi.org/10.1007/s00382-016-3276-3

Kučerová M, Beck C, Philipp A, Huth R (2017) Trends in frequency and persistence of atmospheric circulation types over Europe derived from a multitude of classifications. Int J Climatol 37:2502–2521. https://doi.org/10.1002/joc.4861

Lapp S, Byrne J, Kienzle S, Townshend I (2002) Linking global circulation model synoptics and precipitation for western North America. Int J Climatol 22:1807–1817. https://doi.org/10.1002/joc.851

Lorenzo MN, Ramos AM, Taboada JJ, Gimeno L (2011) Changes in present and future circulation types frequency in northwest Iberian Peninsula. PLoS ONE 6:e16201. https://doi.org/10.1371/journal.pone.0016201

Lund IA (1963) Map-pattern classification by statistical methods. J Appl Meteorol 2:56–65

Lupikasza E (2010) Relationships between occurrence of high precipitation and atmospheric circulation in Poland using different classifications of circulation types. Phys Chem Earth 35:448–455. https://doi.org/10.1016/j.pce.2009.11.012

Lynch A, Uotila P, Cassano JJ (2006) Changes in synoptic weather patterns in the polar regions in the twentieth and twenty-first centuries, part 2: Antarctic. Int J Climatol 26:1181–1199. https://doi.org/10.1002/joc.1306

McKendry IG, Steyn DG, McBean G (1995) Validation of synoptic circulation patterns simulated by the Canadian climate centre general circulation model for western North America. Atmos-Ocean 33:809–825. https://doi.org/10.1080/07055900.1995.9649554

McKendry IG, Stahl K, Moore RD (2006) Synoptic sea-level pressure patterns generated by a general circulation model: comparison with types derived from NCEP/NCAR re-analysis and implications for downscaling. Int J Climatol 26:1727–1736. https://doi.org/10.1002/joc.1337

Meehl GA, Covey C, Taylor KE, Delworth T, Stouffer RJ, Latif M, McAvaney B, Mitchell JFB (2007) The WCRP CMIP3 multimodel dataset: a new era in climate change research. Bull Am Meteorol Soc 88:1383–1394. https://doi.org/10.1175/BAMS-88-9-1383

Palm V, Sepp M, Truu J, Ward RD, Leito A (2017) The effect of atmospheric circulation on spring arrival of short- and long-distance migratory bird species in Estonia. Boreal Env Res 22:97–114

Pastor MA, Casado MJ (2012) Use of circulation types classifications to evaluate AR4 climate models over the Euro-Atlantic region. Clim Dyn 39:2059–2077. https://doi.org/10.1007/s00382-012-1449-2

Perez J, Menendez M, Mendez FJ, Losada IJ (2014) Evaluating the performance of CMIP3 and CMIP5 global climate models over the north-east Atlantic region. Clim Dyn 43:2663–2680. https://doi.org/10.1007/s00382-014-2078-8

Pfahl S (2014) Characterising the relationship between weather extremes in Europe and synoptic circulation features. Nat Hazards Earth Syst Sci 14:1461–1475. https://doi.org/10.5194/nhess-14-1461-2014

Philipp A, Della-Marta PM, Jacobeit J, Fereday DR, Jones PD, Moberg A, Wanner H (2007) Long-term variability of daily North Atlantic–European pressure patterns since 1850 classified by simulated annealing clustering. J Clim 20:4065–4095. https://doi.org/10.1175/JCLI4175.1

Philipp A, Bartholy J, Beck C, Erpicum M, Esteban P, Fettweis R, Huth R, James P, Jourdain S, Kreienkamp F, Krennert T, Lykoudis S, Michaelides S, Pianko K, Post P, Rasilla Álvarez D, Schiemann R, Spekat A, Tymvios FS (2010) Cost733cat—a database of weather and circulation type classifications. Phys Chem Earth 35:360–373. https://doi.org/10.1016/j.pce.2009.12.010

Philipp A, Beck C, Huth R, Jacobeit J (2016) Development and comparison of circulation type classifications using the COST 733 dataset and software. Int J Climatol 36:2671–2809. https://doi.org/10.1002/joc.3920

Plavcová E, Kyselý J (2012) Atmospheric circulation in regional climate models over Central Europe: links to surface air temperature and the influence of driving data. Clim Dyn 39:1681–1695. https://doi.org/10.1007/s00382-011-1278-8

Plavcová E, Kyselý J (2013) Projected evolution of circulation types and their temperatures over Central Europe in climate models. Theor Appl Climatol 114:625–634. https://doi.org/10.1007/s00704-013-0874-4

Poli P et al (2016) ERA-20C: An atmospheric reanalysis of the twentieth century. J Clim 29:4083–4097. https://doi.org/10.1175/JCLI-D-15-0556.1

Rohrer M, Croci-Maspoli M, Appenzeller C (2017) Climate change and circulation types in the Alpine region. Meteorol Z 26:83–92. https://doi.org/10.1127/metz/2016/0681

Rust HW, Vrac M, Lengaigne M, Sultan B (2010) Quantifying differences in circulation patterns based on probabilistic models: IPCC AR4 multimodel comparison for the North Atlantic. J Clim 23:6573–6589. https://doi.org/10.1175/2010JCLI3432.1

Schiemann R, Frei C (2010) How to quantify the resolution of surface: an, example for alpine precipitation. Phys Chem Earth 35:403–410. https://doi.org/10.1016/j.pce.2009.09.005

Schoof JT, Pryor SC (2006) An evaluation of two GCMs: simulation of North American teleconnection indices and synoptic phenomena. Int J Climatol 26:267–282. https://doi.org/10.1002/joc.1242

Shepherd TG (2014) Atmospheric circulation as a source of uncertainty in climate change projections. Nat Geosci 7:703–708. https://doi.org/10.1038/ngeo2253

Sheridan SC, Lee CC (2011) The self-organizing map in synoptic climatological research. Prog Phys Geogr 35:109–119. https://doi.org/10.1177/0309133310397582

Stefan S, Necula C, Georgescu F (2010) Analysis of long-range transport of particulate matters in connection with air circulation over Central and Eastern part of Europe. Phys Chem Earth 35:523–529. https://doi.org/10.1016/j.pce.2009.12.008

Stryhal J, Huth R (2017) Classifications of winter Euro-Atlantic circulation patterns: an intercomparison of five atmospheric reanalyses. J Clim 30:7847–7861. https://doi.org/10.1175/JCLI-D-17-0059.1

Stryhal J, Huth R (2018) Trends in winter circulation over the British Isles and central Europe in twenty-first century projections by 25 CMIP5 GCMs. Clim Dyn. https://doi.org/10.1007/s00382-018-4178-3

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Tveito OE (2010) An assessment of circulation type classifications for precipitation distribution in Norway. Phys Chem Earth 35:395–402. https://doi.org/10.1016/j.pce.2010.03.044

Tveito OE, Huth R (2016) Circulation-type classifications in Europe: results of the COST 733 Action. Int J Climatol 36:2671–2672. https://doi.org/10.1002/joc.4768

Uppala SM et al (2005) The ERA-40 re-analysis. Q J R Meteorol Soc 131(612):2961–3012. https://doi.org/10.1256/qj.04.176

Ustrnul Z, Czekierda D, Wypych A (2010) Extreme values of air temperature in Poland according to different atmospheric

circulation classifications. Phys Chem Earth 35:429–436. https://doi.org/10.1016/j.pce.2009.12.012

Valverde V, Pay MT, Baldasano JM (2015) Circulation-type classification derived on a climatic basis to study air quality dynamics over the Iberian Peninsula. Int J Climatol 35:2877–2897. https://doi.org/10.1002/joc.4179

van Ulden AP, van Oldenborgh GJ (2006) Large-scale atmospheric circulation biases and changes in global climate model simulations and their importance for climate change in Central Europe. Atmos Chem Phys 6:863–881. https://doi.org/10.5194/acp-6-863-2006

Vial J, Osborn TJ (2012) Assessment of atmosphere-ocean general circulation model simulations of winter northern hemisphere atmospheric blocking. Clim Dyn 39:95–112. https://doi.org/10.1007/s00382-011-1177-z

Wójcik R (2015) Reliability of CMIP5 GCM simulations in reproducing atmospheric circulation over Europe and the North Atlantic: a statistical downscaling perspective. Int J Climatol 35:714–732. https://doi.org/10.1002/joc.4015

Wood JL, Harrison S, Turkington TAR, Reinhardt L (2016) Landslides and synoptic weather trends in the European Alps. Clim Change 136:297–308. https://doi.org/10.1007/s10584-016-1623-3

Zappa G, Shaffrey LC, Hodges KI (2013) The ability of CMIP5 models to simulate North Atlantic extratropical cyclones. J Clim 26:5379–5396. https://doi.org/10.1175/JCLI-D-12-00501.1