LETTER

Reductions in daily continental-scale atmospheric circulation biases between generations of global climate models: CMIP5 to CMIP6

Alex J Cannon
Climate Research Division, Environment and Climate Change Canada, Victoria, British Columbia, Canada
E-mail: alex.cannon@canada.ca

Keywords: climate model, atmospheric circulation, model evaluation, regional climate, global climate

Abstract
This study evaluates and compares historical simulations of daily sea-level pressure circulation types over 6 continental-scale regions (North America, South America, Europe, Africa, East Asia, and Australasia) by 15 pairs of global climate models from modeling centers that contributed to both Coupled Model Intercomparison Project Phase 5 (CMIP5) and CMIP6. Atmospheric circulation classifications are constructed using two different methodologies applied to two reanalyses. Substantial improvements in performance, taking into account internal variability, are found between CMIP5 and CMIP6 for both frequency (24% reduction in global error) and persistence (12% reduction) of circulation types. Improvements between generations are robust to different methodological choices and reference datasets. A modest relationship between model resolution and skill is found. While there is large intra-ensemble spread in performance, the best performing models from CMIP6 exhibit levels of skill close to those from the reanalyses. In general, the latest generation of climate models should provide less biased simulations for use in regional dynamical and statistical downscaling efforts than previous generations.

1. Introduction
The day-to-day variability in surface weather is controlled by the large-scale atmospheric circulation. The ability of climate models to correctly simulate the chaotic, dynamical aspects of climate, which exert themselves regionally, is much more difficult to assess than thermodynamic aspects, which can be assessed readily through globally integrated quantities (Shepherd 2019). Specific features of the large-scale circulation, for example storm tracks (Yang et al 2018) and atmospheric blocking (Davini and D’Andrea 2016) over regions of the globe, have been used to diagnose climate model biases in atmospheric circulation. More generally, regional biases in relevant circulation features have been assessed with the aid of circulation pattern or weather type classifications—automated classification schemes that group daily or sub-daily circulation fields into a small number of dominant modes of high-frequency variability (e.g. McKendry et al 1995, 2006, Schoof and Pryor 2006, Otero et al 2017, Stryhal and Huth 2018). These classifications are region-specific, identifying the main circulation patterns that control weather conditions in the region of interest.

Correctly simulating the historical atmospheric circulation in global climate models (GCMs) is relevant for constructing regional climate change scenarios. Dynamical downscaling and statistical post-processing methods rely on credible input data from GCMs, i.e. ‘garbage in, garbage out’ (Hall 2014). For example, Prein et al (2019) found that the driving GCM had a large effect on the ability of regional climate models (RCMs) contributing to the North American Coordinated Regional Downscaling Experiment (CORDEX) (Gutowski et al 2016) to represent historical weather types. RCM biases stemming from GCM biases persisted into future climate change simulations. In an assessment of statistical post-processing methods, Maraun et al (2017) concluded that bias correction can lead to implausible future climate change signals when applied over regions with substantial biases in atmospheric circulation. Both Prein et al (2019) and Maraun et al (2017) advocate for screening of GCMs used to
develop regional climate change assessments based on their ability to simulate the large-scale atmospheric circulation. A related area of research—regional emergent constraints—is based on the identification of strong empirical relationships between historical model biases and future climate change signals at regional scales. Biases in atmospheric circulation have been identified as regional constraints on precipitation change in North America, Europe, and Australia (Simpson et al 2015, Grose et al 2017, Zhang and Soden 2019).

Given how important the representation of atmospheric circulation is in being able to make credible projections of regional climate change, GCMs should be evaluated in terms of their historical circulation biases. With the recent completion of simulations contributing to the Coupled Model Intercomparison Project Phase 6 (CMIP6) (Eyring et al 2016), the main goal of this study is to assess whether there have been substantial reductions—relative to internal variability—in circulation biases between CMIP5 and CMIP6. This is quantified over the two generations as a whole and in terms of successive contributions from each modeling center, including an assessment of dependence on model resolution. Further, GCM performance is framed in terms of observational uncertainty, as measured by differences between reanalyses, to determine if models are converging toward a level of skill comparable to observationally-constrained datasets.

Achieving this goal is, however, hampered by the region-specific nature of the circulation features that control weather in different parts of the globe. Global-scale analyses are usually confined to assessing a particular aspect of the atmospheric circulation, for example surface storm tracks (Booth et al 2017), whereas regional-scale analyses that target multiple circulation features, for example those based on circulation pattern classifications, are usually only conducted for a particular region, for example Europe (Otero et al 2017, Stryhal and Huth 2018) or North America (Prein et al 2019). From a methodological standpoint, the current study bridges these two evaluation philosophies by assessing atmospheric circulation biases in GCMs over 6 continental-scale regions—roughly corresponding to major CORDEX domains—covering the inhabited terrestrial areas of the globe. This is done by 1) applying automated circulation classification schemes, in a consistent fashion, to daily sea-level pressure (SLP) data from reanalyses over each region; 2) classifying GCM-simulated historical SLP fields based on the observationally-based schemes; and 3) assessing CMIP5 and CMIP6 climate model performance in terms of the consistency between the resulting simulated and observed frequency and persistence of each circulation type. Methodological uncertainty is probed through the use of multiple circulation classification schemes and data processing choices; observational uncertainty through the use of different reanalyses; and internal variability with the use of outputs from a large initial condition climate model ensemble. The end result is the first global comparison of regional GCM circulation biases between CMIP5 and CMIP6 models.

2. Data

Observed and simulated historical atmospheric circulation is characterized in terms of daily mean SLP fields over the 6 continental-scale regions shown in figure 1. Two reanalyses included in the Collaborative Reanalysis Technical Environment (Potter et al 2018) are used as observational reference datasets: the NOAA-CIRES Twentieth Century Reanalysis (20CRv2c) (2° grid spacing) (Compo et al 2011) and the JMA JRA-55 (~55-km horizontal resolution) (Kobayashi et al 2015). The two reanalyses assimilate different types of observations. 20CRv2c only assimilates surface pressure observations and prescribes the distribution of sea surface temperature and sea ice, whereas JRA-55 assimilates three dimensional data from surface, radiosonde, and satellite observations. The fundamental differences in resolution, assimilation strategies, and ultimately observational constraints means that differences between these two datasets likely provide a liberal estimate of observational uncertainty.

Both reanalyses have sufficiently long periods of overlapping record with the historical CMIP5 (ending in 2005) and CMIP6 simulations (ending in 2014). For simplicity, a common 45 year period (1960–2004) shared by the reanalyses, CMIP5 simulations, and CMIP6 simulations is used for evaluation. While this spans the pre-/post-satellite era (starting in 1979), a long period of record is important to minimize sampling errors associated with natural variability (Shepherd 2014). Based on Wilcoxon signed-rank tests, significant shifts in pre-/post-satellite era time series of intra-annual means and standard deviations of SLP are identified in 2 of the 6 regions (AFR and SAM), but these are common to both 20CRv2c, which does not assimilate satellite observations, and JRA-55, which does.

Daily mean SLP fields are obtained for the CMIP5 and CMIP6 GCMs listed in table 1. To facilitate the inter-generational comparison, GCM outputs are limited to modeling centers that contributed to both CMIP5 and CMIP6. In total, 15 CMIP5 GCMs and 15 CMIP6 GCMs, from 14 modeling centers, are included in the assessment. As shown in table 1, horizontal resolution varies widely between the GCMs and, as noted before, the reanalyses. As the focus here is on the large-scale atmospheric circulation, all data are conservatively remapped onto a regular 2.5° grid, which is close to that of the coarsest resolution GCMs, before analysis. The same approach was taken by Sillmann et al (2013) in an evaluation of climate extremes indices in CMIP5 models. With the
exception of the large 50 member initial condition ensemble of CanESM5 simulations (Swart et al 2019), which is used to estimate internal variability, outputs from a single ensemble member are used for model evaluation.

Finally, the atmospheric circulation classification schemes selected for use in this analysis rely on observations of surface weather conditions to guide the partitioning of the daily SLP fields into discrete types. Daily mean surface air temperature, wind speed, and precipitation fields over land areas within each region, again remapped onto a 2.5° grid, are obtained for 20CRv2c and JRA-55.

3. Circulation classifications

The atmospheric circulation classification schemes used here group, in an automated fashion, time series of daily SLP fields over each region into a set of $k$ discrete types; each day is assigned to exactly one type. Results can be communicated as a daily time series of a categorical variable with $k$ categories—or, equivalently, $k$ time series of binary indicator variables, one for each atmospheric circulation type—and a composite SLP field constructed from the daily fields assigned to each of the $k$ types. Fundamentally, the primary goal is the same as any clustering algorithm, namely for the SLP fields within a type to be as similar to one another as possible while simultaneously being as different as possible from the SLP fields assigned to the other types. Once the classification schemes have been trained on observations, new SLP data, e.g. from a GCM or another reanalysis, can be assigned to one of the existing $k$ types, often by measuring the statistical distance of a new field to the composite SLP fields and assigning to the nearest type.

Figure 1. Locations of the 6 continental-scale regions: North America (NAM), Europe (EUR), East Asia (EAS), South America (SAM), Africa (AFR), and Australasia (AUS).

Table 1. CMIP5 and CMIP6 models and their nominal diagonal grid spacings (in ° at the equator). Modeling center contributions with no change in horizontal resolution between generations are indicated by an asterisk.

| Modeling center | Model | CMIP5 Grid (°) | CMIP6 Grid (°) |
|-----------------|-------|----------------|----------------|
| Beijing Climate Center | bcc-csm1-1-m | 2.81 | BCC-CSM2-MR | 1.13 |
| Canadian Centre for Climate Modelling and Analysis | CanESM2 | 2.81 | CanESM5 | 2.81* |
| National Center for Atmospheric Research | CESM1-CAM5 | 1.10 | CESM2 | 1.10* |
| Centre National de Recherches Météorologiques and Centre Européen de Recherche et de Formation Avancée | CNRM-CM5 | 1.41 | CNRM-CM6-1 | 1.41* |
| EC-Earth-Consortium | EC-EARTH | 1.12 | EC-Earth3 | 0.70 |
| LASG/IAP, Beijing, China | FGOALS-g2 | 2.81 | FGOALS-g3 | 2.13 |
| Geophysical Fluid Dynamics Laboratory | GFDL-CM3 | 2.26 | GFDL-CM4 | 1.00 |
| Geophysical Fluid Dynamics Laboratory | GFDL-ESM2M | 2.26 | GFDL-ESM4 | 1.00 |
| Met Office Hadley Centre and Natural Environment Research Council (CMIP6) | HadGEM2-ES | 1.59 | UKESM1-0-LL | 1.59* |
| Institute for Numerical Mathematics, Russian Academy of Sciences | inmcm4 | 1.77 | INM-CM5-0 | 1.77* |
| Institut Pierre Simon Laplace | IPSL-CM5A-LR | 2.96 | IPSL-CM6A-LR | 1.98 |
| JAMSTEC, AORI, NIES, and R-CCS | MIROC5 | 1.41 | MIROC6 | 1.41* |
| Max Planck Institute for Meteorology | MPI-ESM-MR | 1.88 | MPI-ESM1-2-HR | 0.94 |
| Meteorological Research Institute | MRI-ESM1 | 1.13 | MRI-ESM2-0 | 1.13* |
| Norwegian Climate Center | NorESM1-M | 2.21 | NorESM2-LM | 2.21* |
A comprehensive review of atmospheric circulation classification schemes, most attempting to meet the aforementioned goal of within-type homogeneity/between-type heterogeneity, can be found in Huth et al (2008). Practical examples can be found in the Cost733cat catalogue and software for Europe (Philipp et al 2010, Demuzere et al 2010). However, because the evaluation here is motivated by the desire to inform regional climate change studies, a secondary goal, the ability to discriminate surface weather conditions, is also taken into consideration by focusing on guided or supervised classification schemes. In this case, a secondary set of surface weather variables—here terrestrial surface air temperature, wind speed, and precipitation observed coincident with SLP—provide ‘inductive bias’ to the clustering algorithms that form the core of the classification schemes.

To account, in part, for methodological uncertainty, two guided atmospheric circulation classification schemes are applied separately to observational data from each reanalysis. The clustered canonical correlation analysis (CCCA) follows the regression-guided clustering approach of Cannon (2012a) and the multivariate regression tree (MRT) scheme follows Cannon et al (2002) and Cannon (2012b). In CCCA, SLP (acting as predictor variables) and surface weather fields (acting as predictand variables) are preprocessed using the Barnett and Preisendorfer (1987) principal component analysis (PCA) truncation approach to CCA (see also Livezey and Smith 1999); retained left (SLP) canonical variates are then entered into a $k$-means clustering algorithm, which identifies the atmospheric circulation types. In MRT, truncated SLP PC scores and individual SLP grid cell data are used as predictor variables in a multivariate regression tree—a binary decision tree machine learning algorithm that recursively partitions the predictor data to best cluster the predictand data—with truncated surface weather PCs serving as predictand variables; atmospheric circulation types are defined by the terminal nodes reached in the decision tree. For sake of brevity, readers are referred to the original references for more details on the underlying algorithms.

Under the conceptual categorization proposed by Philipp et al (2010), the CCCA and MRT schemes are both ‘optimization’ methods, with MRT also being a ‘threshold based’ method with rules for assignment that are optimized rather than predetermined. As described by Demuzere et al (2010), the use of surface weather variables to guide the CCCA and MRT classifications is similar in spirit to the optional Cost733cat conditioning of SLP-based circulation types towards a target variable.

A consistent approach to variable preprocessing and application of the atmospheric circulation classification schemes is used in all continental-scale regions. In each, separate classifications are developed for extended boreal cold (ONDJFM) and warm (AMJJAS) seasons. S-mode PCA is applied to the correlation matrix of observed gridded data; variables are weighted to account for latitudinal differences in grid area. SLP PCs are truncated to explain 90% of variance in the original gridded SLP data. Surface weather fields are transformed (fourth root for precipitation, square root for wind speed, and untransformed for temperature) to be more normally distributed prior to combined PCA. Surface weather PCs are truncated to explain 50% of variance in the original data. For CCCA, canonical variates with canonical correlations in excess of 0.45 ($\sim 20\%$ explained variance) are retained for entry into the $k$-means clustering algorithm. Cluster analysis is repeated 50 times, each time starting with random centers, with the best optimization used for the final classification.

### 4. Experimental setup

Atmospheric circulation classifications are developed for each of the 6 regions using the two classification schemes (CCCA and MRT) and two observationally-constrained datasets (JRA-55 and 20CRv2c). The number of circulation types per extended season $k$ is set to 16, which is consistent with the upper end of the range included in the Cost733cat database over smaller European domains (Philipp et al 2010). To assess the sensitivity of results to $k$, a separate set of classifications is developed with $k$ set to 8.

Once developed, SLP fields from the historical GCM simulations are entered into the atmospheric circulation classifications. For a given region, extended season, and GCM, each day is assigned to one of the $k$ observed types. In practice, this means first projecting the simulated SLP fields onto the PCs and, for CCCA, canonical variates, and then assigning to the existing CCCA and MRT types. For CCCA, this is done by calculating the Euclidean distance between the GCM SLP canonical variates on a given day and the observed composite mean SLP canonical variates for each of the $k$ types, and then assigning to the type with the minimum distance. For MRT, the GCM SLP PC scores and individual SLP grid values on a given day are entered into the binary decision tree; the terminal node reached determines the type.

As pointed out by Demuzere et al (2009) and discussed further by Stryhal and Huth (2018), the manner in which GCM SLP data are preprocessed can explain a large portion of apparent skill in...
reproducing observed atmospheric circulation features. For example, it is common practice in statistical downscaling applications to remove GCM biases in grid cell means and variances prior to projecting onto observed PCs (Schubert 1998). If the same practice is adopted when evaluating GCMs in terms of subsequent classification performance, then these biases will no longer contribute to biases in atmospheric circulation and one will be left with an overly optimistic assessment of skill. Any such processing steps, e.g. removal of grid cell or domain mean biases, will artificially improve apparent GCM performance.

In this study, all classifications are run, in turn, using three preprocessing strategies, each more aggressive than the previous at removing SLP biases (table 2). In the first, ‘domain centered’, biases in SLP (i.e. over an entire region) are removed for each GCM. In part, this is deemed necessary to minimize potential differences in results between CMIP5 and CMIP6 generation GCMs due to changes in historical forcing datasets and differences in climate sensitivity. Overall, domain mean biases are small in both CMIP5 and CMIP6; absolute magnitudes range from 0.3 hPa averaged over CMIP5 GCMs in summer for SAM to 1.1 hPa over CMIP6 GCMs in winter for AFR. There is no evidence of systematic improvements in domain mean bias between model generations—considering all combinations of GCM, reference reanalysis, region, and extended season, CMIP6 GCMs are less biased than their CMIP5 counterparts in 46% of cases. In the second strategy, ‘grid cell centered’, grid cell mean biases are removed. In the third, ‘grid cell scaled’, biases in both grid cell means and variances are removed. Unless noted otherwise, subsequent results are reported based on domain mean bias removal.

5. Evaluation measures

GCM performance is evaluated by comparing observed and simulated frequencies of occurrence and day-to-day persistence of the atmospheric circulation types. The bias in simulated frequency of occurrence $f_{s,i}$ of the $i$th of $k$ circulation types in season $s$, relative to the observed frequency $o_{s,i}$, is assessed using the logarithmic accuracy ratio

$$LAR_{s,i} = \log_{10} \left( \frac{f_{s,i}}{o_{s,i}} \right)$$

which is a symmetrical measure of relative error (i.e. weather types that are simulated to occur 1/10th
and 10 times as frequently as observed yield LAR values of -1 and 1 respectively. Bias in simulated persistence for the $i$th weather type is given by the run length error

$$RLE_{s,i} = \bar{f}_{s,i} - \bar{o}_{s,i},$$  

which is simply the difference in the mean number of consecutive days that a weather type occurs in simulations $\bar{f}_{s,i}$ and observations $\bar{o}_{s,i}$.

Aggregate measures of frequency performance for each region are summarized in terms of the median absolute LAR

$$\text{MALAR} = \text{median}\{ |LAR_{\text{cold},1}|, \ldots, |LAR_{\text{cold},k}|, \ldots, |LAR_{\text{warm},1}|, \ldots, |LAR_{\text{warm},k}| \}$$  

and persistence performance by the mean absolute RLE

$$\text{MARLE} = \frac{1}{2k} \sum_{s \in \{\text{cold,warm}\}} \sum_{i=1}^{k} |RLE_{s,i}|.$$  

For sake of readability, values of MALAR are transformed into positive relative frequency errors

$$\text{MALAR'} = 10^{\text{MALAR}} - 1.$$  

Accuracy ratios of $f/o = 1.2$ and $1/1.2$ then both yield a MALAR’ value of 0.2, i.e. a positive relative frequency error of 20%.

Values of the evaluation measures are calculated for the 30 GCMs listed in table 1 for each combination of experimental settings listed in table 2. In addition, each reanalysis is treated as a pseudo GCM and its errors with respect to the classifications developed using the other reanalysis are also calculated. The mean inter-reanalysis error is used as an estimate of observational uncertainty. Finally, errors are also calculated for 50 members of a large initial condition ensemble of CanESM5 historical simulations (Swart et al 2019). The standard deviation of errors across the CanESM5 large ensemble (CanESM5 LE) serves as an estimate of internal variability.

6. Regional results

Frequency and persistence errors are summarized in figure 2 as portrait plots with the 30 GCMs arranged in columns and the 6 continental-scale regions in rows. For reference, inter-reanalysis errors (a measure of observational uncertainty) are reported as a separate column, as are standard deviations (scaled by a factor of 10) of CanESM5 LE errors (a measure
Figure 4. Substantial differences in (a) frequency and (b) persistence performance between paired CMIP5 and CMIP6 contributions from the same modeling center. Substantial differences are defined as those that exceed 2 standard deviations of internal variability as estimated from CanESM5 LE.

of internal variability). Overall, differences between the CMIP5 and CMIP6 ensembles are masked by the inherent differences in dominant atmospheric circulation processes operating in each region, as well as differences in observational uncertainty between regions. In general, there have been modest improvements in regional performance between generations of models. Ensemble mean reductions in frequency error from CMIP5 to CMIP6 range from 6% (AFR) to 39% (SAM) and in persistence error from 5% (EUR) to 22% (SAM). Visually, however, it is evident that the best performing models, especially in CMIP6, exhibit levels of error that approach those between the 20CRv2c and JRA-55 reanalyses (cf GFDL-CM4 and Reanalyses columns).

To more clearly visualize the differences between model generations, figure 3 shows ranks of GCM performance—pooling CMIP5 and CMIP6 models—from best to worst. Here, the separation between CMIP5 and CMIP6 becomes more apparent. While there is large inter-model spread in both CMIP5 and CMIP6 ensembles, a small number of GCMs perform well across the majority of regions in terms of both frequency and persistence errors (median rank < 10); all are from the CMIP6 ensemble (UKESM1-0-LL, GFDL-CM4, GFDL-ESM4, MPI-ESM1-2-HR, EC-Earth3, and MIROC6).

Evaluating a single ensemble member from each GCM means that differences in skill between model generations may be due simply to internal variability rather than true changes in model performance. To get a rough sense for whether differences between CMIP5 and CMIP6 contributions from a given modeling center are substantial, figure 4 indicates
region/reanalysis/classification scheme combinations with changes in skill whose magnitude exceeds two standard deviations of the CanESM5 LE internal variability estimate. For frequency errors, CMIP6 models offer a substantial improvement over the corresponding CMIP5 model in 41% of cases, versus a substantial deterioration in just 18% of cases. For persistence errors, an improvement is seen in 49% of cases and deterioration in 13% of cases. When looking at GCMs from each modeling center, increased skill is seen in more regions than decreased skill in 80% of CMIP5/CMIP6 model pairs for frequency and 93% of model pairs for persistence.

7. Global results

Results reported in the previous section focused on regional performance. Figure 5 switches focus and instead shows changes in performance of individual models averaged globally over the regions, both in terms of ranks and errors relative to observational uncertainty. The underlying error measures used to calculate model ranks are mean values of frequency error and persistence error taken over the 6 regions, and, to show error relative to observational uncertainty, median values of scaled frequency error and scaled persistence error over the regions. Global results are, as expected, consistent with those reported in previous figures. While there is a substantial overlap in performance between the CMIP5 and CMIP6 ensembles, the majority of modeling centers improved atmospheric circulation performance in their most recent models. No model jumps in rank across more than two quartiles, although the IPSL contribution moves from the lowest ranked GCM in CMIP5 to just outside the top quartile in both frequency and persistence in CMIP6. Globally, there has been a reduction in mean frequency error of 24% and in persistence error of 12% from CMIP5 to CMIP6, although, as seen earlier, this is not distributed evenly across all regions. Further, a small number of CMIP6 models (GFDL-CM4, GFDL-ESM4, and MIROC6) exhibit errors that approach the level of observational uncertainty considered in this study (i.e. both
frequency and persistence errors lie within 1.25 times inter-reanalysis error).

8. Sensitivity to experimental setup

Methodological and observational uncertainty has been incorporated into results shown so far through the use of two atmospheric classification schemes and two reanalyses, in all cases based on $k = 16$ seasonal circulation types and domain mean centering of GCM SLP fields. As shown in table 2, additional experiments have been conducted to explore sensitivity to the number of seasonal circulation types ($k = 8$) and GCM SLP bias removal, both via grid cell centering (mean) and grid cell scaling (mean and variance). With domain mean bias removal and $k = 8$, results are largely consistent, both regionally and globally, to those for $k = 16$. The CMIP5-to-CMIP6 reduction in mean frequency error changes to 21% (versus 24% for $k = 16$); the reduction in persistence error remains unchanged at 12%. Regional reductions in error are also similar, ranging in magnitude from 7% (AFR, 6% for $k = 16$) to 39% (SAM, unchanged) and in persistence error from 4% (AFR, 8% for $k = 16$) to 15% (NAM, 20% for $k = 16$), as are substantial increases and decreases in regional performance (not shown).

As expected (and illustrated in figure 6), global error rates fall substantially, by construction, when GCM grid cell mean biases and grid cell mean/variance biases are removed in data preprocessing. With that said, global improvements in performance are still noted between the CMIP5 and CMIP6 ensembles. For the grid cell centered experiment, there is a 9% reduction in global frequency error and an 8% reduction in persistence error; this remains unchanged (+/- 0.2% points) when removing biases in grid variance.

9. Resolution dependence

In the absence of model outputs from the same GCM run at different resolutions, it is extremely difficult
to diagnose how (and which) improvements to a climate model lead to improvements in atmospheric circulation biases. With that said, 7 out of the 15 model pairs reduced horizontal grid spacing from CMIP5 to CMIP6 (table 1). Figure 6 shows the relationship between grid spacing and global average frequency and persistence errors. Spearman rank correlations with grid spacing for the domain centered experiment equal 0.51 for frequency error and 0.63 for persistence error; 0.64 and 0.64, respectively, for the grid cell centered experiment; and 0.74 and 0.64 for the grid cell scaled experiment. Overall, there is a modest positive association (from ~25% to 55% explained variance) between grid spacing and frequency and persistence errors, depending on how GCM bias is handled. On this latter point, note the large reduction in error that occurs between domain mean bias removal and grid cell mean bias removal (from figure 6(a) to 6(b), and the subsequent minimal change when errors in grid variance are further removed (from figure 6(b) to 6(c)). The dominant component of atmospheric circulation errors is thus likely mean biases at the grid cell level (e.g. location errors), rather than biases in variance.

10. Conclusion

This study evaluates historical simulations of daily sea-level pressure atmospheric circulation types over 6 continental-scale regions (North America, South America, Europe, Africa, East Asia, and Australasia) by 15 pairs of global climate models from modeling centers that contributed to both CMIP5 and CMIP6. It is the first global assessment of improvements in daily circulation performance in CMIP6 models relative to their CMIP5 counterparts. Substantial improvements in performance, taking into account internal variability, are found moving from CMIP5 to CMIP6 for both frequency (22% reduction in error) and persistence (12% reduction) of daily circulation types, with little dependence on methodological choices like observational reference dataset, classification scheme, and number of circulation types. Results do indicate a modest relationship between model resolution and skill. As future work, when outputs from GCMs run in multiple resolutions (e.g. HadGEM3-GC31-LI/MM, MPI-ESM-LR/HR/XR, and NorESM2-LM/MM) are available, they should be evaluated to determine if improvements in performance are attributable to increases in resolution.

Consistent with findings of Demuzere et al (2009) and Stryhal and Huth (2018), the manner in which GCM data are preprocessed can explain a large portion of apparent classification performance. GCM biases that are removed, either explicitly or implicitly, cannot contribute to biases in atmospheric circulation, with the potential for overly optimistic assessments of skill. The bulk of the results here rely on domain mean bias removal, mainly to ensure that results are not contaminated by differences due to changes in forcing specifications between CMIP5 and CMIP6, but also due to differences in climate model sensitivity. There is substantial scope for exploring additional classification schemes, as in Stryhal and Huth (2018), or other observational reference datasets.

The global nature of this study naturally precludes a detailed assessment of physical mechanisms and specific circulation features contributing to model errors in specific regions. Still, the errors do point to certain regions, for example AFR, where improvements in atmospheric circulation performance have slowed or stalled relative to other regions. It is also possible to interrogate saved results from the classification schemes to probe physical mechanisms more deeply. This is left for future work.

While the intra-ensemble spread in performance is still large, the best performing models from CMIP6 have levels of skill that are comparable to the reanalysis skill. The main caveat here is that observational uncertainty is estimated based on two reanalyses with large differences in data assimilation and modeling strategies. Regardless, from a practical standpoint, climate model contributions to CMIP6 should provide less biased atmospheric circulation patterns to subsequent regional dynamical and statistical downscaling efforts than previous generations.

Acknowledgment

We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP5 and CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP activities and ESGF. The data that support the findings of this study are openly available from the ESGF.

Data availability statement

The data that support the findings are openly available at the links listed below. 20CRv2c is available at DOI: 10.5065/D6N877TW and https://www.esrl.noaa.gov/psd/data/20thC_Rean/; JRA-55 at DOI: 10.5065/D6HH6H41 and https://jra.kishou.go.jp/JRA-55/; and CMIP data at https://esgf-node.llnl.gov/.

ORCID iD
Alex J Cannon © https://orcid.org/0000-0002-8025-3790
References

Barnett T P and Preisendorfer R 1987 Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis Mon. Weather Rev. 115 1825–50

Booth J R, Kwon Y O, Ko S, Small R J and Msadek R 2017 Spatial patterns and intensity of the surface storm tracks in CMIP5 models J. Clim. 30 4965–81

Cannon A J 2012a Regression-guided clustering: a semisupervised method for circulation-to-environment synoptic classification J. Appl. Meteorol. Clim. 51 185–90

Cannon A J 2012b Semi-supervised multivariate regression trees: putting the ‘circulation’ back into a ‘circulation-to-environment’ synoptic classifier Int. J. Climatol. 32 2251–4

Cannon A J, Whitfield P H and Lord E R 2002 Synoptic map-pattern classification using recursive partitioning and principal component analysis Mon. Weather Rev. 130 1187–206

Compo G P et al 2011 The twentieth century reanalysis project Q. J. R. Meteorol. Soc. 137 1–28

Davini P and D’Andrea F 2016 Northern hemisphere atmospheric blocking representation in global climate models: twenty years of improvements J. Clim. 29 8823–40

Demuzere M, Kassomenos P and Philipp A 2010 The COST733 circulation type classification software: an example for surface ozone concentrations in central europe Theor. Appl. Climatol. 105 143–66

Demuzere M, Werner M, NPM van L and Roeckner E 2009 An analysis of present and future ECHAM5 pressure fields using a classification of circulation patterns Int. J. Climatol. 29 1796–1810

Eyring V, Bony S, Meehl G A, Senior C A, Stevens B, Stouffer R J and Taylor K E 2016 Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization Geosci. Model Dev. 9 1937–58

Grose M R, Risbey J S, Moise A F, Osbrough S, Heady C, Wilson L and Erwin T 2017 Constraints on southern Australian rainfall change based on atmospheric circulation in CMIP5 simulations J. Clim. 30 225–42

Gutowski Jr W J et al 2016 WCRP coordinated regional downscaling experiment (CORDEX): a diagnostic MIP for CMIP6 Geosci. Model Dev. 9 4087–95

Hall A 2014 Projecting regional change Science 346 1461–2

Huth R, Beck C, Philipp A, Demuzere M, Ustrnul Z, Cahynová M, Kyvelý J and Tweito O E 2008 Classifications of atmospheric circulation patterns Ann. New York Acad. Sci. 1146 105–52

Kobayashi S et al 2015 The JRA-55 reanalysis: general specifications and basic characteristics J. Meteorol. Soc. Japan. Ser. II 93 5–48

Livezey R E and Smith T M 1999 Considerations for use of the Barnett and Preisendorfer (1987) algorithm for canonical correlation analysis of climate variations J. Clim. 12 303–5

Maraun D et al 2017 Towards process-informed bias correction of climate change simulations Nat. Clim. Change 7 764–73

McKendry I G, Stahl K and Moore R D 2006 Synoptic sea-level pressure patterns generated by a general circulation model: comparison with types derived from NCEP/NCAR re-analysis and implications for downsampling Int. J. Climatol. 26 1727–36

McKendry I G, Steyn D G and McBean G 1995 Validation of synoptic circulation patterns simulated by the Canadian Climate Centre General Circulation Model for western North America Atmos.–Ocean 33 809–25

Otero N, Sillmann J and Butler T 2017 Assessment of an extended version of the Jenkinson–Collison classification on CMIP5 models over europe Clim. Dyn. 50 1559–79

Philipp A et al 2010 Cost733cat—a database of weather and circulation type classifications Phys. Chem. Earth, Parts A/B/C 35 360–73

Potter G L, Carriere L, Hertz J, Bosilovich M, Duffy D, Lee T and Williams D N 2018 Enabling reanalysis research using the Collaborative REAnalysis Technical Environment (CREATE) Bull. Am. Meteorol. Soc. 99 677–87

Prin A F, Bukovsky M S, Mearns L O, Bruyère C L and Done J M 2019 Simulating North American weather types with regional climate models Frontiers Environ. Sci. 7 36

Schoof J T and Pryor S C 2006 An evaluation of two GCMs: simulation of North American teleconnection indices and synoptic phenomena Int. J. Climatol. 26 267–82

Schoebert S 1998 Downscaling local extreme temperature changes in south-eastern Australia from the CSIRO mark2 GCM Int. J. Climatol. 18 1419–38

Shepherd T G 2014 Atmospheric circulation as a source of uncertainty in climate change projections Nat. Geosci. 7 703–8

Shepherd T G 2019 Storyline approach to the construction of regional climate change information Proc. R. Soc. A. 475 20190013

Sillmann J, Kharin V V, Zhang X, Zwiers F W and Bronaugh D 2013 Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. model evaluation in the present climate J. Geophys. Res.: Atmos. 118 1716–33

Simpson I R, Seager R, Ting M and Shaw T A 2015 Causes of change in Northern Hemisphere winter meridional winds and regional hydroclimate Nat. Clim. Change 6 65–70

Stryhal J and Huth R 2018 Classifications of winter atmospheric circulation patterns: validation of CMIP5 GCMs over Europe and the north Atlantic Clim. Dyn. 52 3575–98

Swart N C et al 2019 The Canadian Earth System Model version 5 (CanESM5.0.3) Geosci. Model Dev. 12 4823–73

Yang M, Li X, Zuo R, Chen X and Wang L 2018 Climatology and interannual variability of winter north Pacific storm track in CMIP5 models Atmosphere 9 79

Zhang B and Soden B J 2019 Constraining climate model projections of regional precipitation change Geophys. Res. Lett. 46 10522–31