Geophysical Research Letters

The Sensitivity of Euro-Atlantic Regimes to Model Horizontal Resolution

K. Strommen[1] , I. Mavilia[2] , S. Corti[2] , M. Matsueda[1,3] , P. Davini[2] , J. von Hardenberg[2] , P.-L. Vidale[4] , and R. Mizuta[5]

[1]Department of Physics, University of Oxford, Oxford, UK, [2]Istitutio di Scienze dell'Atmosfera e del Clima, Consiglio Nazionale delle Richerche, Rome, Italy, [3]Center for Computational Sciences, University of Tsukuba, Tsukuba, Japan, [4]Department of Meteorology, University of Reading, Reading, UK, [5]Meteorological Research Institute, Tsukuba, Japan

Abstract There is growing evidence that the atmospheric dynamics of the Euro-Atlantic sector during winter is driven in part by the presence of quasi-persistent regimes. However, general circulation models typically struggle to simulate these with, for example, an overly weakly persistent blocking regime. Previous studies have showed that increased horizontal resolution can improve the regime structure of a model but have so far only considered a single model with only one ensemble member at each resolution, leaving open the possibility that this may be either coincidental or model dependent. We show that the improvement in regime structure due to increased resolution is robust across multiple models with multiple ensemble members. However, while the high-resolution models have notably more tightly clustered data, other aspects of the regimes may not necessarily improve and are also subject to a large amount of sampling variability that typically requires at least three ensemble members to surmount.

1. Introduction

Predicting the evolution of the atmospheric state over time can be understood as a question of determining likely trajectories along the atmospheres climate attractor in phase space. Since the 1980s, evidence has begun to accumulate that suggests the geometry of this attractor exhibits interesting local structure, which manifests itself in the form of quasi-persistent weather regimes (e.g., Vautard, 1990 and Michelangeli et al., 1995, and discussion within for early history; more recently, see, e.g., Straus et al., 2007; Straus, 2010; Woollings, Hannachi, Hoskins & Turner 2010; Woollings, Hannachi, & Hoskins 2010; Franzke et al., 2011; and Hannachi et al., 2017). In particular, such regimes have been identified in the Euro-Atlantic region, and there is a growing recognition of their importance in modulating European weather (Ferranti et al., 2015; Frame et al., 2013; Matsueda & Palmer, 2018) and, conjecturally, the regional response to anthropogenic forcing (Corti et al., 1999; Palmer, 1999). Representing these regimes correctly is therefore an important goal for any general circulation model (GCM).

The studies Dawson et al. (2012) and Dawson and Palmer (2015) demonstrated that a GCM's ability to capture Euro-Atlantic regimes appears to depend on the horizontal resolution of the model. In particular, improvements in the spatial structure, geometric robustness (by which we mean the extent to which the data can be divided into tightly knit clusters), and persistence statistics of the regimes were all identified upon increasing the resolution. However, these studies used only one model, with a single ensemble member at each resolution. This leaves open the question as to how robust this resolution dependence is across models, as well as the possibility that sampling variability may be playing a role. In this paper, we address these issues by examining the impact of increasing resolution on three different models, each with three ensemble members. Besides examining the impact of resolution, we also evaluate the impact of using an ensemble: By concatenating multiple ensemble members, we can obtain larger data sets, effectively reducing the impact of excessive noise and/or poorly constrained regimes.

We will show that, for all three models considered, the low-resolution models struggle to replicate the regime structure seen in reanalysis data sets. None of the nine individual simulations achieve comparable levels of clustering to that of reanalysis, and while the regime patterns on average have a relatively high spatial correlation with those of reanalysis, the spread is often large with some individual members performing notably poorly. Persistence of the blocking regime is also systematically underestimated in all simulations,

with the model tending to vacate the regime faster than reanalysis. Increasing the horizontal resolution leads to notably more tightly clustered data, with a few individual high-resolution simulations achieving a regime structure comparable to reanalysis. A systematic improvement in the persistence statistics of the blocking regime is also seen across all the models; no such systematic change is identified for the other regimes. This is consistent with the results of the multimodel study conducted in Schiemann et al. (2017), demonstrating improvements in atmospheric blocking (as measured using more standard European blocking indices) with increased horizontal resolution. However, no systematic improvement in the spatial patterns of the regimes are seen, with the net impact being a slight degradation compared to the low-resolution patterns.

We also show that, for single ensemble members at low resolution, there can be a notable spread around the average value of the metrics in question, suggesting that sampling variability for these quantities can be large. For the low-resolution models, one generally needs to use all three members in order to generate regime statistics comparable to reanalysis over the approximately 30-year periods considered, while for high resolution, two ensemble members suffice. This supports the idea that models at lower resolution have too weak regime structure and that increased resolution can be expected to ameliorate this to some extent. It also highlights the fact that, for models with weaker regime structure, a large sample size of simulation years is necessary to diagnose regimes robustly.

2. Data and Methods

2.1. Data

We use model data from three models, all run in atmosphere-only mode, covering between 25 and 31 years in the period 1979–2011. Each model produced three simulations at both a "low" and "high" resolution, where the exact meaning of low and high varies between the models. This leaves us with nine low-resolution simulations and nine high-resolution simulations to compare across: due to the varying nature of the resolution increase, we always group results by model, to see the relative impact in each model. We also note that the monikers "low" and "high" resolution are essentially arbitrary here and are used simply for convenience to denote the lower/higher of the two available resolutions, rather than any objective measure. The study is therefore only concerned with the effect of *increasing* resolution, and not on the exact impact of any *particular* resolution choice.

The first model is EC-Earth v3.1, an Earth system model maintained by the EC-Earth Consortium (Hazeleger et al., 2012). Its atmospheric component is based on the Integrated Forecasting System model cycle 36r4, developed by the European Centre for Medium-Range Weather Forecasts. The integrations were made as part of the Climate SPHINX Project (Davini et al., 2017) and covered the period 1979–2008. Ten such ensemble members were produced in the SPHINX Project: Only three were considered in the analysis in order to make the comparison across all models and resolutions as uniform as possible. Results were found to be qualitatively identical irrespective of which three ensemble members were selected, so the results presented here used the first three members. The low-resolution simulations had a spectral truncation of TL255, or roughly 80-km grid spacing near the equator. The high-resolution simulations had a spectral truncation of TL511, corresponding to around 40-km grid spacing. Both use 91 levels in the vertical. Note that the studies Dawson et al. (2012) and Dawson and Palmer (2015) considered an earlier version of the same model.

The second model is the U.K. Met Office model HadGEM3-GA3 (Walters et al., 2011). The integrations were run as part of the UPSCALE project (Mizielinski et al., 2014) and cover the period 1986–2011. The low-resolution simulations were performed on a N216 grid, corresponding to roughly 60-km grid spacing near the equator, while the high-resolution simulations were done on a N512 grid, corresponding to roughly 25-km grid spacing. Both configurations use 85 levels in the vertical.

Finally, we used the Japanese Meterorological Research Institute (MRI) model AGCM3.2 (Mizuta et al., 2012). The low-resolution version was integrated at TL95 resolution, corresponding to roughly 180-km grid spacing, while the high-resolution simulations were integrated at TL319 resolution, corresponding to roughly 60-km grid spacing. Both use 64 levels in the vertical. The three ensemble members cover the period 1979–2010.

The primary reanalysis product used to act as our reference data set was the European Centre for Medium-Range Weather Forecasts dataset ERA-Interim (Dee et al., 2011), covering the period 1979–2011. To bolster confidence in the results, and to estimate the potential sampling variability of the metrics inherent to the real atmosphere, we also utilized the NCEP/NCAR reanalysis data set (Kalnay et al., 1996), hereafter

referred to as NCEP, and the Japanese 55-year reanalysis (abbr. JRA55; see Kobayashi et al., 2015). In general, the difference between the three data sets for a given metric were small, being an order of magnitude smaller than the model biases and the impacts seen from a resolution change. Because all three data sets show such close agreement, we will only present explicit values for ERA-Interim and NCEP in tables/figures.

All data sets were first interpolated down to a common 2.5° regular grid prior to carrying out computations.

It is important to note that the actual range of resolutions considered in the paper are relatively narrow, leaving open the question as to whether the observed changes could be expected for *any* change in resolution. Also important to note is that none of the high-resolution models were tuned separately from the low resolution: The simulations therefore differ only in the resolution itself.

2.2. Methodology

Regimes are defined using a *k*-means clustering algorithm, following the method in Straus et al. (2007) (see also Michelangeli et al., 1995): A regime thereby corresponds to something resembling a fixed point in phase space, around which observed atmospheric states tend to cluster. The algorithm is applied to the daily geopotential height field at 500 hPa, considered over a Euro-Atlantic domain defined by 30°–90°N, 80°W–40°E. We then restrict the data to the December-January-February winter period for each available year. A climatological cycle is obtained from this field and smoothed with a 5-day running mean; this smoothed cycle is then removed from the original field to produce a time series of daily geopotential height anomalies. In order to make the algorithm tractable, the dimensionality of the field is reduced using an empirical-orthogonal-function (EOF) decomposition. Only the first four (unnormalized) EOFs are retained: These explain more than 50% of the variance for both models and reanalysis. It was found that using more EOFs, explaining up to 80% of the variance, produced quantitatively similar results. Restricting to the first four alone also focuses the analysis on the large-scale patterns, which we are interested in. The *k*-means clustering algorithm applied to this final field will then produce clusters that maximize the following *optimal ratio*:

$$Optimal\ Ratio = \frac{Intercluster\ variance}{Intracluster\ variance}, \tag{1}$$

where the intercluster variance refers to the variance between the cluster centroids (weighted by the number of points in each cluster) and the intracluster variance refers to the average variance of the differences between the cluster centroids and the data points associated to that cluster. A large intercluster variance therefore implies that the centroids are well separated from each other, while a small intracluster variance implies the points of each cluster are located close to their respective centroid. A large optimal ratio is therefore associated with a more clearly robust regime structure.

The presence of high autocorrelation in the data can influence the *k*-means clustering algorithm, potentially inflating the optimal ratio. This is exacerbated by the fact that the algorithm will always generate the number of clusters one asks for, meaning large optimal ratios may occur purely by chance. This raises two issues. First, how does one evaluate the statistical significance of the regimes generated? Second, given the influence of autocorrelation, which may vary between simulations, how does one compare optimal ratios across multiple models? Both these questions are addressed by defining a "significance" metric in the following manner. A statistical null hypothesis is assumed in which the phase space in question has no particular regime structure and therefore the atmosphere has no preferred locations or directions of movement in phase space. Concretely, the null hypothesis posits that the phase space is equivalent to that expected from assuming that each of the four coordinates of the atmosphere, in this truncated four-dimensional phase space, are behaving like independent Markov processes with a fixed mean, variance, and lag-1 correlation equaling those of the data set in question. The assumption of the process being first order (and therefore using the lag-1 correlation) was justified by plotting the autocorrelation of our data sets and noting that these were very well captured by a basic exponential decay. While skewness in the data can, in principle, also influence clustering, we found little sensitivity to the computed metrics when adding skewness to the null hypothesis, and this was therefore ignored. Randomly generating four such Markov processes defines new atmospheric coordinates, which will populate the four-dimensional phase space; by applying the clustering algorithm to this data set, one computes the optimal ratio for the clusters produced. Repeating this for 500 different synthetic data sets generates a distribution of optimal ratios that could be expected from our null hypothesis. We define the *sharpness* of the clustering to be the percentage of points in this distribution *below* the optimal ratio actually computed with the original data set. A sharpness close to 100% therefore implies
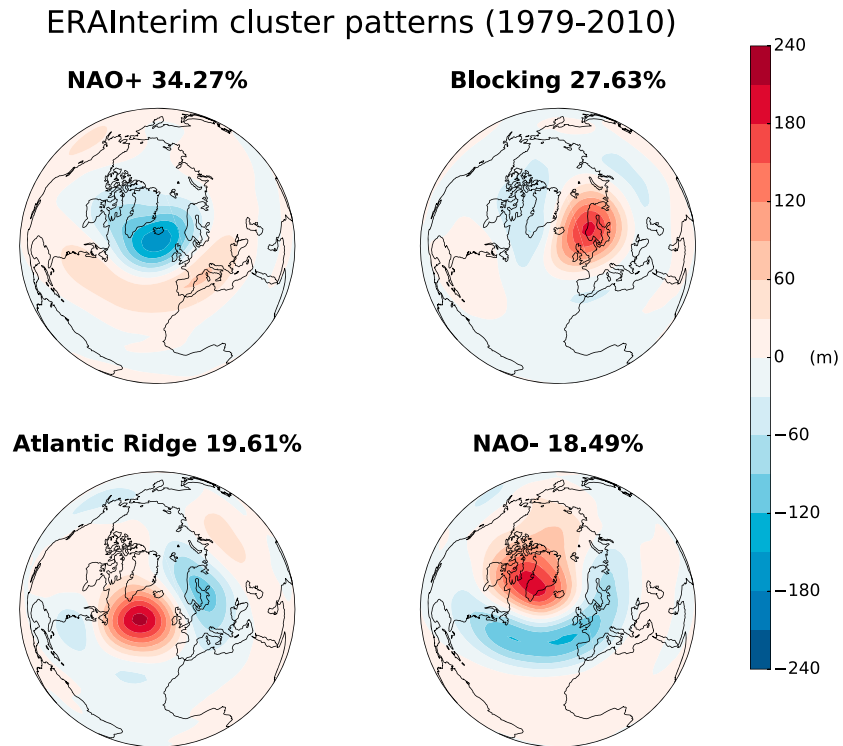
Figure 1. Spatial patterns of the four regimes defined by the cluster centroids for ERA-Interim (1979–2010). Obtained by applying *k*-means clustering to the geopotential height anomalies at 500 hPa, restricted to the Euro-Atlantic region. The percentages indicate the frequency of occurrence of that regime during the entire time period.

a large optimal ratio unlikely to have arisen by chance from an atmosphere with no regime structure. Crucially, by effectively "normalizing" the optimal ratio relative to the autocorrelation of the underlying data, the sharpness metric can readily be compared across multiple models. In Dawson et al. (2012), this metric was referred to simply as "significance," but as we are using this as a metric in and of itself, we rename it to avoid potential confusion. However, it is important to keep in mind that this is a measure of confidence in clustering and so may change according to the sample size.

Applying the algorithm to reanalysis data shows, as noted in Dawson et al. (2012), that four clusters are the minimum number required to produce statistically significant regimes (i.e., a sharpness exceeding 95%). For this reason, we restrict our attention to a four-regime picture, both for reanalysis and our model data. Figure 1 shows the spatial patterns of these four regimes in the reanalysis data, which agree well with previous studies. These patterns are generated by taking the mean across all the daily fields that are sorted into a given cluster by the algorithm.

We will focus on three metrics for assessing the models representation of regimes. First, the sharpness metric will be used as a measure of the geometric robustness of regimes. A high sharpness suggests strongly defined regimes but can be obtained with regimes that do not look like those in reality. Therefore, second, pattern correlation between the regime patterns of the models and those in reanalysis is used as a measure of how similar to reanalysis the diagnosed regimes are. Finally, we look at the level of day-to-day regime persistence.

Note that EOFs are always computed independently for each data set in question. In particular, when multiple ensemble members are concatenated, EOFs are computed for the concatenated data set.

3. Results
3.1. Regime Sharpness
Figure 2 shows the impact of both increased resolution and increased ensemble size. In plot (a), the blue (red) dots/crosses/triangles associated with the ECEall/HadGEMall/MRIall labels are sharpness metrics for the relevant low-resolution (high-resolution) model obtained after concatenating all three ensemble members, effectively tripling the sample size compared to the reanalysis data sets. The horizontal black line
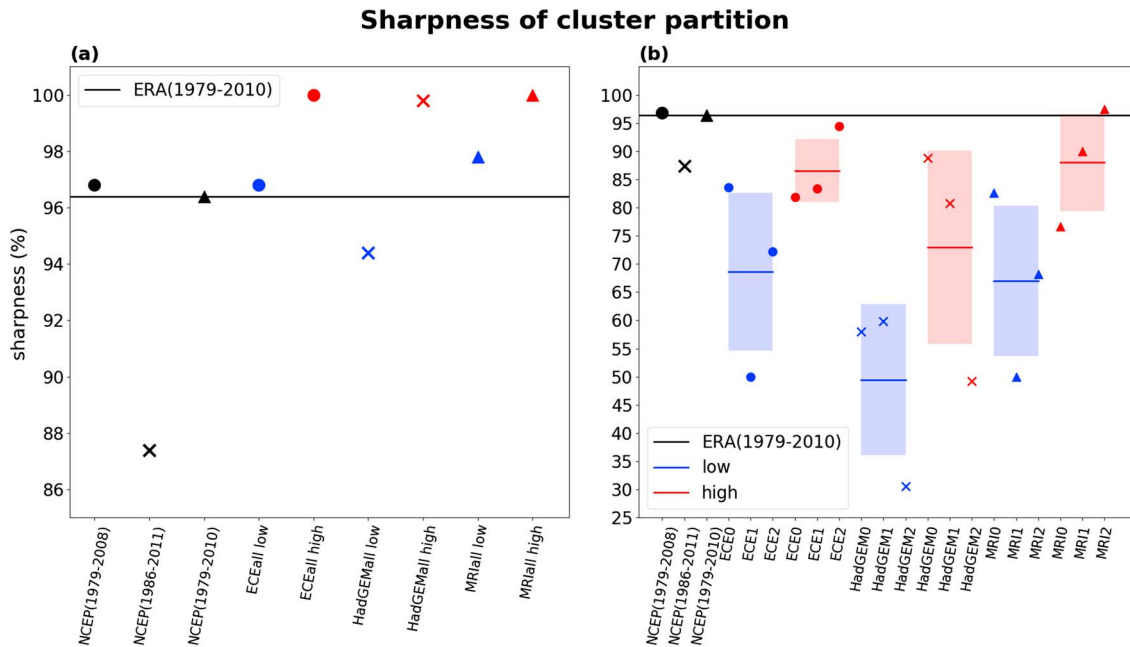
Sharpness of cluster partition

Figure 2. The sharpness metric for reanalysis products and the three models. In (a), for NCEP reanalysis over the three different periods considered (black dots, crosses, and triangles), along with that computed for all low/high EC-Earth resolution data sets concatenated (ECEall low/high, blue/red dots), all low/high HadGEM data sets concatenated (HadGEM low/high, blue/red crosses), and all low/high MRI data sets concatenated (MRIall low/high, blue/red triangles). In (b), sharpness for the same reanalysis data (black dots, crosses, and triangles), and the sharpness metric for each individual ensemble member: EC-Earth low/high resolution (blue/red dots), HadGEM low/high resolution (blue/red crosses), and MRI low/high resolution (blue/red triangles). The mean and standard deviation are indicated by the horizontal colored line and shading, respectively. The horizontal black line in both plots shows the sharpness of ERA-Interim over the period 1979–2010.

shows the sharpness of ERA-Interim over the period 1979–2010, while the black circles/crosses/triangles show the sharpness metrics for the NCEP reanalysis computed over the different time periods covered by the three models. Note that the sharpness of NCEP during the period 1979–2010 matches that of ERA-Interim almost exactly, suggesting that this metric is well constrained. The two reanalysis data sets also produce nearly identical sharpness research metrics when viewed over the two other time periods. In plot (b), we show the sharpness metrics for each individual ensemble member of the three models at low resolution (blue dots/crosses/triangles) and high resolution (red dots/crosses/triangles). The blue/red lines in these triplets indicate the mean of the three points, and the shading encloses one standard deviation.

The four sharpness metrics for reanalysis are almost identical (approximately 97%), with the exception of that covering the period 1986–2011, where the metric is notably lower (approximately 87%). As the time period covered is shorter, it is possible that this is simply random sampling variability in a significantly clustered system. It is also possible that the extent to which regime dynamics drive the atmosphere is non-stationary, with the period 1979–1985 being particularly tightly clustered. However, since the concatenation of all three MRI experiments covering 1986–2011 have a sharpness close to 100%, we will assume that the drop in sharpness seen with NCEP is sampling variability. Therefore, a difference in sharpness of up to 10% might be expected by chance alone.

The impact of increased resolution is apparent for all three models, where sharpness increases whether looking at individual members or after concatenating all three. It can also be seen that increasing the sample size, by using more than one ensemble member, increases sharpness. This can be understood by noting that a model with weak regime structure will be more prone to producing clusters that cannot be robustly distinguished from random noise when using a small sample size. Increasing the sample size effectively filters out noise in phase space. This can be examined further by evaluating the average sharpness metric obtained after concatenating two random members from each simulation. The results are shown in Figure 3, which summarizes the dependence of sharpness on ensemble size (i.e., sample size). It can be seen that the high-resolution models typically need twice the sample size of reanalysis data to achieve comparable regime structure, while for low resolution three times the sample size may still not suffice.
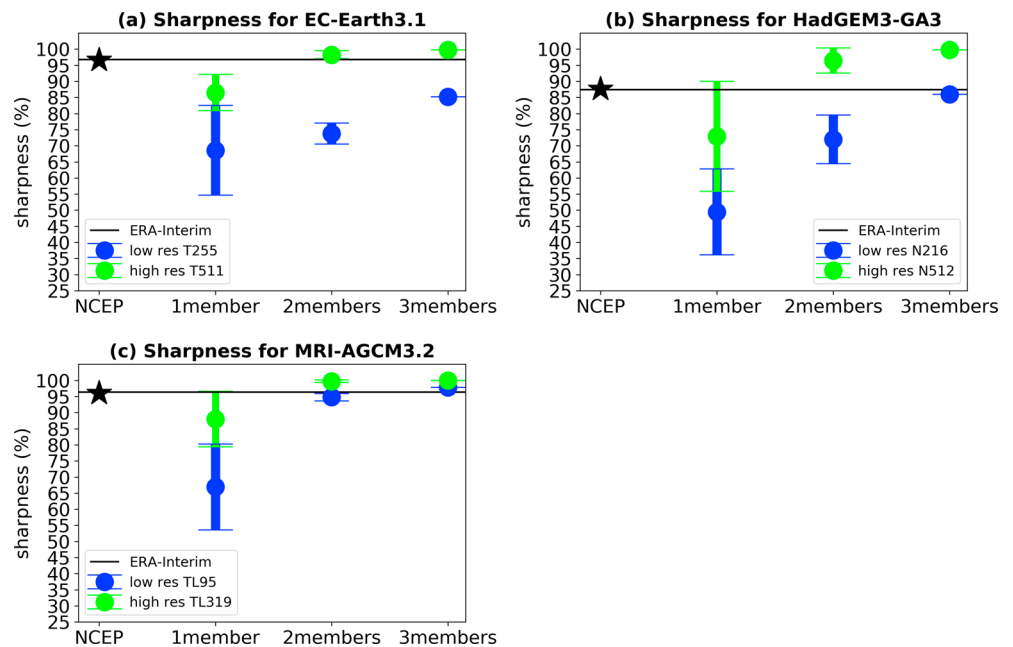
Figure 3. Dependence of sharpness metric on number of ensemble members for (a) EC-Earth3.1, (b) HadGEM3-GA3, and (c) MRI-AGCM3.2. In each, values for "1 member" is the average sharpness across the three low-resolution (high-resolution) members, "2 members" the average sharpness when concatenating combinations of two ensemble members (over all such combinations), and "3 members" the sharpness obtained after concatenating all three members: low resolution in blue and high resolution in green. Error bars show one standard deviation around the mean. The horizontal black line shows the sharpness of ERA-Interim, and the black star shows the value of NCEP, over the relevant time periods.

We note also (see Figure 2) that the increase in sharpness upon increasing the resolution is roughly twice as large as the maximum difference in sharpness between the two reanalysis products ERA-Interim and NCEP, with the former approximately 20% and the latter at 10%. This, combined with the fact that the increase is consistent across all three models, suggests that this improvement is statistically robust.

3.2. Regime Locations

Table 1 shows the mean pattern correlation of the model clusters (across the three ensemble members) relative to ERA-Interim, with the error given as twice the standard deviation as computed across the three entries. High pattern correlation implies the cluster centroids of the model are in approximately the same location in phase space as reanalysis and is therefore a measure of the regimes being in the correct location or not. In particular, the regime patterns will look similar to those in Figure 1 when this correlation is high. The contents of the table demonstrate that there is significant variability in this quantity, with no clear improvement across all three models upon increasing the resolution.

While for the Atlantic Ridge and NAO− regimes, the spread in the pattern correlation goes down when increasing resolution; for NAO+ and Blocking, it frequently goes up. The mean correlation itself also shows no systematic improvement and is sometimes degraded when increasing resolution. This can be seen more starkly in the "All" quantities in the table, which show the pattern correlations obtained from the model regimes after concatenating all three ensemble members. While we saw in the previous section that after combining the data in this way, the sharpness notably increased with resolution, the pattern correlation tends to slightly *decrease*, with the average decrease across all models and regimes being approximately −0.05. The average change in pattern correlation across all unconcatenated experiments and all regimes, is approximately −0.007, with a standard deviation of 0.08, indicating no significant change. Therefore, the impact is at best neutral, and more likely a slight degradation.

These results suggest first that there remains considerable uncertainty in any model estimate of these quantities as diagnosed from even three ensemble members and, second, that one cannot expect to see a consistent improvement in pattern correlation upon increasing resolution: The opposite may in fact be expected to happen. It is unclear to what extent sampling error is influencing the estimates: It is possible that with a much

Table 1
Mean Pattern Correlation of the Regimes Across the Three Simulations and Errors (Given as Two Standard Deviations)

	NAO+	Blocking	Atlantic Ridge	NAO−
ECE-Low	0.96 (±0.02)	0.84 (±0.2)	0.64 (±0.63)	0.87 (±0.2)
ECE-Hi	**0.81** (±0.3)	**0.78** (±0.34)	**0.49** (±0.43)	**0.90** (±0.07)
ECE-Low (all)	0.96	0.90	0.85	0.97
ECE-Hi (all)	**0.99**	**0.96**	**0.81**	**0.90**
HadGEM-Low	0.96 (±0.01)	0.92 (±0.03)	0.93 (±0.19)	0.83 (±0.38)
HadGEM-Hi	**0.93** (±0.04)	**0.89** (±0.04)	**0.91** (±0.06)	**0.80** (±0.20)
HadGem-Low (All)	0.97	0.96	0.98	0.99
HadGem-Hi (All)	**0.94**	**0.89**	**0.95**	**0.79**
MRI-Low	0.74 (±0.56)	0.80 (±0.13)	0.71 ± 0.54	0.88 (±0.23)
MRI-Hi	**0.89** (±0.2)	**0.75** (±0.35)	**0.74** (±0.33)	**0.95** (±0.04)
MRI-Low (All)	0.94	0.84	0.96	0.96
MRI-Hi (All)	**0.93**	**0.78**	**0.81**	**0.96**

Note. Quantities labeled "All" are correlations of the model regimes obtained by concatenating all three simulations into one large data set. Values from high-resolution simulations are highlighted with bold to aid readability.

larger ensemble, one would be able to spot more robust changes. Using all 10 EC-Earth members makes the uncertainty estimates smaller and much more uniform across the four regimes. However, the reduced uncertainty, of ±0.16 on average, still does not allow a statistically significant distinction between high and low resolution for all four regimes.

3.3. Regime Persistence

To assess the impact on persistence, we considered the seasonal persistence probabilities of the four regimes. These are computed for a given season by estimating the probability that if the atmosphere is in a regime on day N, it will remain in the same regime on day $N + 1$. This gives an indication of how persistent that regime tended to be in said season. For each model and each regime, distributions of these seasonal persistence probabilities were computed on the concatenation of the three ensemble members. These distributions were used to assess persistence.

All the models display biases in persistence statistics, but, with the exception of the blocking regime, we observed no systematic such bias, nor did we find a systematic improvement with resolution across all the models. For the blocking regime, all the models underestimate persistence, placing too much weight in the tail of short-lived events. This is consistent with previous studies showing that GCM's tend to underestimate persistent blocking (e.g., Anstey et al., 2013; D'Andrea et al., 1998; Davini & D'Andrea, 2016; Jung et al., 2012; Masato et al., 2013; and Matsueda et al., 2009). The move to higher resolution results in all cases in the distribution shifting closer to reanalysis, implying that levels of persistence have increased for all three models. Indeed, the average persistence across the three models increases by around 1.5%, and the number of occurrences of less persistent years (probabilities lower than 80%) decreases by around 11%. This is in agreement with the recent multimodel study conducted in (Schiemann et al., 2017), which also showed an improvement in blocking persistence upon increasing the resolution. Note that in all these papers atmospheric blocking was defined in a very different way, so our results are *consistent* rather than equivalent. A figure showing this result can be found in the supporting information.

4. Discussion and Conclusions

All three models exhibit deficiencies in the geometric robustness of the regimes (measured with sharpness), the location of the regime centers (how similar the regime patterns are to those in reanalysis), and the persistence lifetime of regimes. Increasing the horizontal resolution leads to a notable improvement in the sharpness metric, suggesting that improvements in certain aspects of atmospheric dynamics are resulting in the atmosphere traversing phase space in a more tightly clustered fashion. Furthermore, the persistence statistics of the blocking regime did systematically improve, consistent with previous studies. However, the

persistence statistics of the other three regimes, as well as the spatial pattern of all four regimes (including blocking), did not systematically improve, with some of the regimes deviating even further from reanalysis in some of the high-resolution simulations. This suggests that, while the increased horizontal resolution in these models is helping some crucial processes responsible for clustering behavior, it fails to alleviate other biases.

Several studies (e.g., Davini & D'Andrea, 2016; Doblas-Reyes et al., 1998; Hinton et al., 2009; Scaife et al., 2010; Woollings, Charlton-Perez, et al., 2010) have implicated biases in the climatological mean state of models to errors in blocking statistics, and Masato et al. (2009) demonstrated the importance of the mean flow in determining the precise locations of blocking events. Given the synoptic scale of the Euro-Atlantic regimes, it is plausible that mean state biases may be important not just for the blocking regime but for all four of the Euro-Atlantic regimes, and these biases are not always improved upon increasing horizontal resolution. Similarly, interannual variability due to global teleconnections such as ENSO, known to influence the NAO (see, e.g., Li & Lau, 2012), may change in nonobvious ways with resolution. Thus, while certain aspects of the atmospheric dynamics may be improving from the increased resolution, leading to more pronounced clustering, the exact locations of these clusters may still be subject to errors due to other biases. This may be exacerbated in situations where the higher-resolution model has not been tuned to achieve as realistic a mean state as the lower-resolution version (as was the case with the models considered here). A detailed examination of this speculation is, however, beyond the scope of the present study.

The above discussion becomes more pertinent when observing that horizontal resolution cannot by itself account for regime skill. Indeed, Figure 2 shows that the low-resolution (180 km) MRI simulations perform almost as well as the significantly higher-resolution (25 km) HadGEM simulations. This implies that other aspects of the model, such as the model mean or small-scale variability related to differing physics parametrisations, are likely equally important for producing robust regimes. Interestingly, MRI also has the coarsest vertical resolution of the three models, seemingly ruling this out as an important factor.

Our results also show that the sampling variability in both the sharpness and spatial correlation metrics is large for both high- and low-resolution experiments. This may be due to excessive noise in the model atmosphere, or the weak regime structure allowing the model to populate phase space too liberally: In a potential well picture of the regimes, if the wells are too shallow, the model atmosphere will not stay trapped for longer periods of time, leading to less tightly clustered data. Either way, increasing the sample size by concatenating data from multiple ensemble members can help filter out this noise. We find that the low-resolution models appear to need three times as much data to detect a comparable regime structure as that found in reanalysis data, while the high-resolution models, by improving the regime structure, need only twice as much data. This implies that multiple ensemble members, or a sufficiently long simulation period, are crucial for a statistically meaningful assessment of regime metrics, with changes due to resolution or other model upgrades being potentially completely invisible due to random noise.

References

Anstey, J. A., Davini, P., Gray, L. J., Woollings, T. J., Butchart, N., Cagnazzo, C., et al. (2013). Multi-model analysis of Northern Hemisphere winter blocking: Model biases and the role of resolution. *Journal of Geophysical Research: Atmospheres*, *118*, 3956–3971. https://doi.org/10.1002/jgrd.50231

Corti, S., Molteni, F., & Palmer, T. N. (1999). Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature*, *398*(6730), 799–802. https://doi.org/10.1038/19745

D'Andrea, F., Tibaldi, S., Blackburn, M., Boer, G., Déqué, M., Dix, M. R., et al. (1998). Northern Hemisphere atmospheric blocking as simulated by 15 atmospheric general circulation models in the period 1979-1988. *Climate Dynamics*, *14*(6), 385–407. https://doi.org/10.1007/s003820050230

Davini, P., & D'Andrea, F. (2016). Northern Hemisphere atmospheric blocking representation in global climate models: Twenty years of improvements? *Journal of Climate*, *29*(24), 8823–8840. https://doi.org/10.1175/JCLI-D-16-0242.1

Davini, P., Von hardenberg, J., Corti, S., Christensen, H. M., Juricke, S., Subramanian, A., et al. (2017). Climate SPHINX: Evaluating the impact of resolution and stochastic physics parameterisations in the EC-Earth global climate model. *Geoscientific Model Development*, *3*, 1383–1402. https://doi.org/10.5194/gmd-10-1383-2017

Dawson, A., & Palmer, T. N. (2015). Simulating weather regimes: Impact of model resolution and stochastic parameterization. *Climate Dynamics*, *44*(7-8), 2177–2193. https://doi.org/10.1007/s00382-014-2238-x

Dawson, A., Palmer, T. N., & Corti, S. (2012). Simulating regime structures in weather and climate prediction models. *Geophysical Research Letters*, *39*, L2180. https://doi.org/10.1029/2012GL053284

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Doblas-Reyes, F. J., Déqué, M., Valero, F., & Stephenson, D. B. (1998). North Atlantic wintertime intraseasonal variability and its sensitivity to GCM horizontal resolution. *Tellus, Series A: Dynamic Meteorology and Oceanography*, *50*(5), 573–595. https://doi.org/10.3402/tellusa.v50i5.14560

Ferranti, L., Corti, S., & Janousek, M. (2015). Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Journal of the Royal Meteorological Society*, *141*(688), 916–924. https://doi.org/10.1002/qj.2411

Frame, T. H., Methven, J., Gray, S. L., & Ambaum, M. H. (2013). Flow-dependent predictability of the North Atlantic jet. *Geophysical Research Letters*, *40*, 2411–2416. https://doi.org/10.1002/grl.50454

Franzke, C., Woollings, T., & Martius, O. (2011). Persistent circulation regimes and preferred regime transitions in the north atlantic. *Journal of the Atmospheric Sciences*, *68*(12), 2809–2825. https://doi.org/10.1175/JAS-D-11-046.1

Hannachi, A., Straus, D. M., Franzke, C. L., Corti, S., & Woollings, T. (2017). *Low-frequency nonlinearity and regime behavior in the Northern Hemisphere extratropical atmosphere* (Vol. 55, pp. 199–234). https://doi.org/10.1002/2015RG000509

Hazeleger, W., Wang, X., Severijns, C., Stefanescu, S., Bintanja, R., Sterl, A., et al. (2012). EC-Earth V2.2: Description and validation of a new seamless Earth system prediction model. *Climate Dynamics*, *39*(11), 2611–2629. https://doi.org/0.1007/s00382-011-1228-5

Hinton, T. J., Hoskins, B. J., & Martin, G. M. (2009). The influence of tropical sea surface temperatures and precipitation on north Pacific atmospheric blocking. *Climate Dynamics*, *33*(4), 549–563. https://doi.org/10.1007/s00382-009-0542-7

Jung, T., Miller, M. J., Palmer, T. N., Towers, P., Wedi, N., Achuthavarier, D., et al. (2012). High-resolution global climate simulations with the ECMWF model in Project Athena: Experimental design, model climate, and seasonal forecast skill. *Journal of Climate*, *25*(9), 3155–3172. https://doi.org/10.1175/JCLI-D-11-00265.1

Kalnay, E., Kanamitsu, M, Kistler, R., Deaven, D., Gandin, M., Iredell, S., et al. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, *77*(i), 437–471. https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., et al. (2015). The JRA-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan, Ser. II*, *93*(1), 5–48. https://doi.org/10.2151/jmsj.2015-001

Li, Y., & Lau, N. C. (2012). Impact of ENSO on the atmospheric variability over the North Atlantic in late Winter-Role of transient eddies. *Journal of Climate*, *25*(1), 320–342. https://doi.org/10.1175/JCLI-D-11-00037.1

Masato, G., Hoskins, B. J., & Woollings, T. J. (2009). Can the frequency of blocking be described by a red noise process? *Journal of the Atmospheric Sciences*, *66*(7), 2143–2149. https://doi.org/10.1175/2008JAS2907.1

Masato, G., Hoskins, B. J., & Woollings, T. (2013). Winter and summer Northern Hemisphere blocking in CMIP5 models. *Journal of Climate*, *26*(18), 7044–7059. https://doi.org/10.1175/JCLI-D-12-00466.1

Matsueda, M., Mizuta, R., & Kusunoki, S. (2009). Future change in wintertime atmospheric blocking simulated using a 20-km-mesh atmospheric global circulation model. *Journal of Geophysical Research*, *114*, D12114. https://doi.org/10.1029/2009JD011919

Matsueda, M., & Palmer, T. (2018). *Estimates of flow-dependent predictability of wintertime Euro-Atlantic weather regimes in medium-range forecasts* (Vol. 144, pp. 1012–1027). https://doi.org/10.1002/qj.3265

Michelangeli, P.-A., Vautard, R., & Legras, B. (1995). Weather regimes: Recurrence and quasi stationarity. *Journal of the Atmospheric Sciences*, *52*(8), 1237–1256. https://doi.org/10.1175/1520-0469(1995)052<1237:WRRAQS>2.0.CO;2

Mizielinski, M. S., Roberts, M. J., Vidale, P. L., Schiemann, R., Demory, M. E., Strachan, J., et al. (2014). High-Resolution Global Climate Modelling: The UPSCALE project, a large-simulation campaign. *Geoscientific Model Development*, *7*, 1629–1640. https://doi.org/10.5194/gmd-7-1629-2014

Mizuta, R. H., Yoshimura, H., Murakami, M., Matsueda, H., Endo, T., Ose, K., et al. (2012). Climate simulations using MRI-AGCM 3.2 with 20-km grid. *Journal of the Meteorological Society of Japan*, *90A*, 233–258.

Palmer, T. N. (1999). A nonlinear dynamical perspective on climate prediction. *Journal of Climate*, *12*(2), 575–591. https://doi.org/10.1175/1520-0442(1999)012<0575:ANDPOC>2.0.CO;2

Scaife, A. A., Woollings, T., Knight, J., Martin, G., & Hinton, T. (2010). Atmospheric blocking and mean biases in climate models. *Journal of Climate*, *23*(23), 6143–6152. https://doi.org/10.1175/2010JCLI3728.1.

Schiemann, R., Demory, M. E., Shaffrey, L. C., Strachana, J., Vidale, P. L., Mizielinski, M. S., et al. (2017). The resolution sensitivity of Northern Hemisphere blocking in four 25-km atmospheric global circulation models. *Journal of Climate*, *30*(1), 337–358. https://doi.org/10.1175/JCLI-D-16-0100.1

Straus, D. M. (2010). Synoptic-Eddy Feedbacks and circulation regime analysis. *Monthly Weather Review*, *138*(11), 4026–4034. https://doi.org/10.1175/2010MWR3333.1

Straus, D. M., Corti, S., & Molteni, F. (2007). Circulation regimes: Chaotic variability versus SST-forced predictability. *Journal of Climate*, *20*(10), 2251–2272. https://doi.org/10.1175/JCLI4070.1

Vautard, R. (1990). Multiple weather regimes over the North Atlantic: Analysis of Precursors and Successors. *Monthly Weather Review*, *118*, 2056–2081. https://doi.org/10.1175/1520-0493(1990)118<2056:MWROTN>2.0.CO;2

Walters, D. N., Best, M. J., Bushell, A. C., Copsey, D., Edwards, J. M., Falloon, P. D., et al. (2011). The Met Office Unified Model Global Atmosphere 3.0/3.1 and JULES Global Land 3.0/3.1 configurations. *Geoscientific Model Development*, *4*(4), 919–941. https://doi.org/10.5194/gmd-4-919-2011

Woollings, T., Charlton-Perez, A., Ineson, S., Marshall, A. G., & Masato, G. (2010). Associations between stratospheric variability and tropospheric blocking. *Journal of Geophysical Research*, *115*, D06108. https://doi.org/10.1029/2009JD012742

Woollings, T., Hannachi, A., & Hoskins, B. (2010). Variability of the North Atlantic eddy-driven jet stream. *Quarterly Journal of the Royal Meteorological Society*, *136*(649), 856–868. https://doi.org/10.1002/qj.625

Woollings, T., Hannachi, A., Hoskins, B., & Turner, A. (2010). A regime view of the North Atlantic oscillation and its response to anthropogenic forcing. *Journal of Climate*, *23*(6), 1291–1307. https://doi.org/10.1175/2009JCLI3087.1