Below we answer comments of the Reviewer 1.
Original comments of the Reviewer 1 are listed in black, our answers - in blue colour.

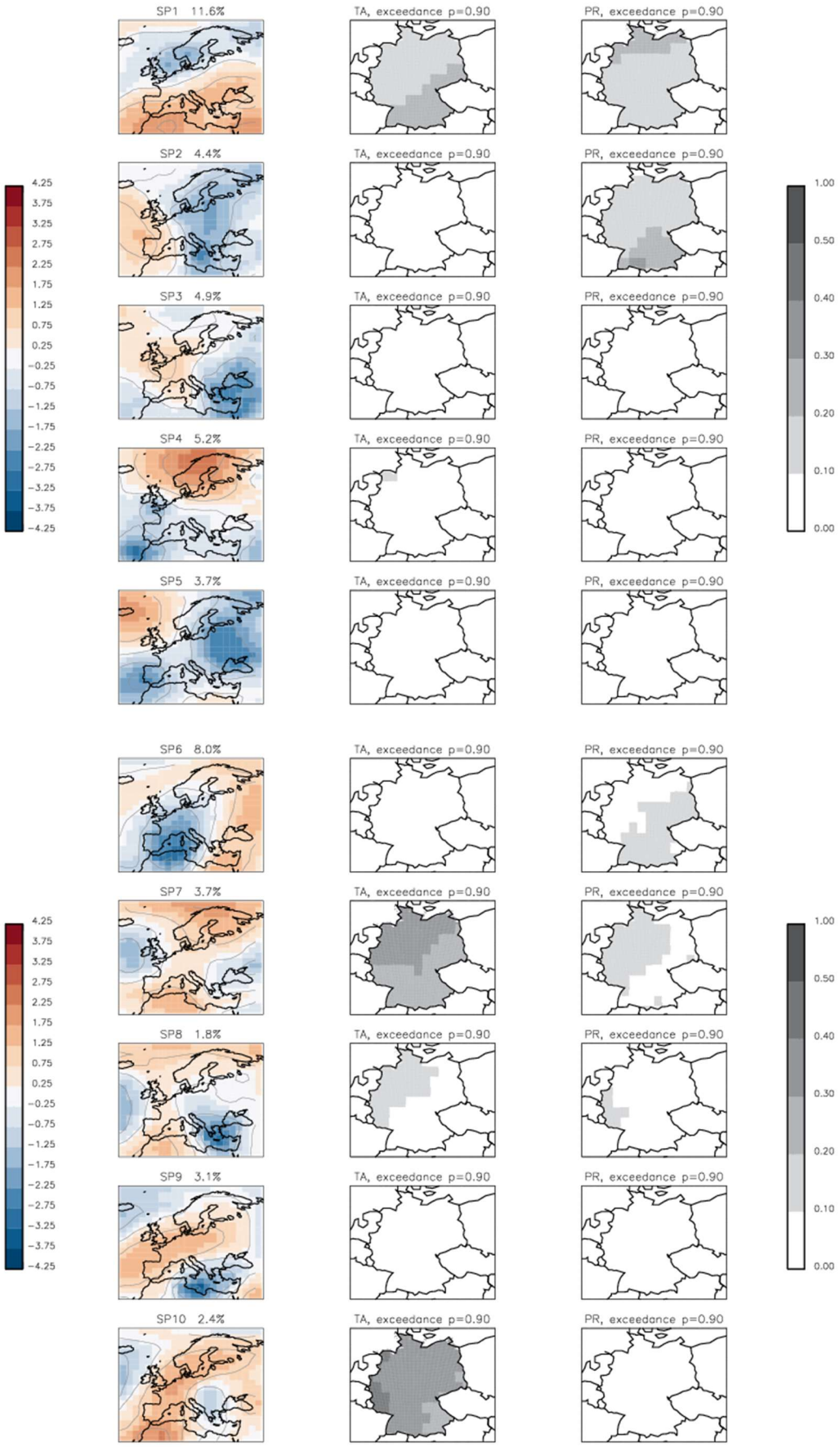
The usefulness of the suggested approach for capturing extreme events

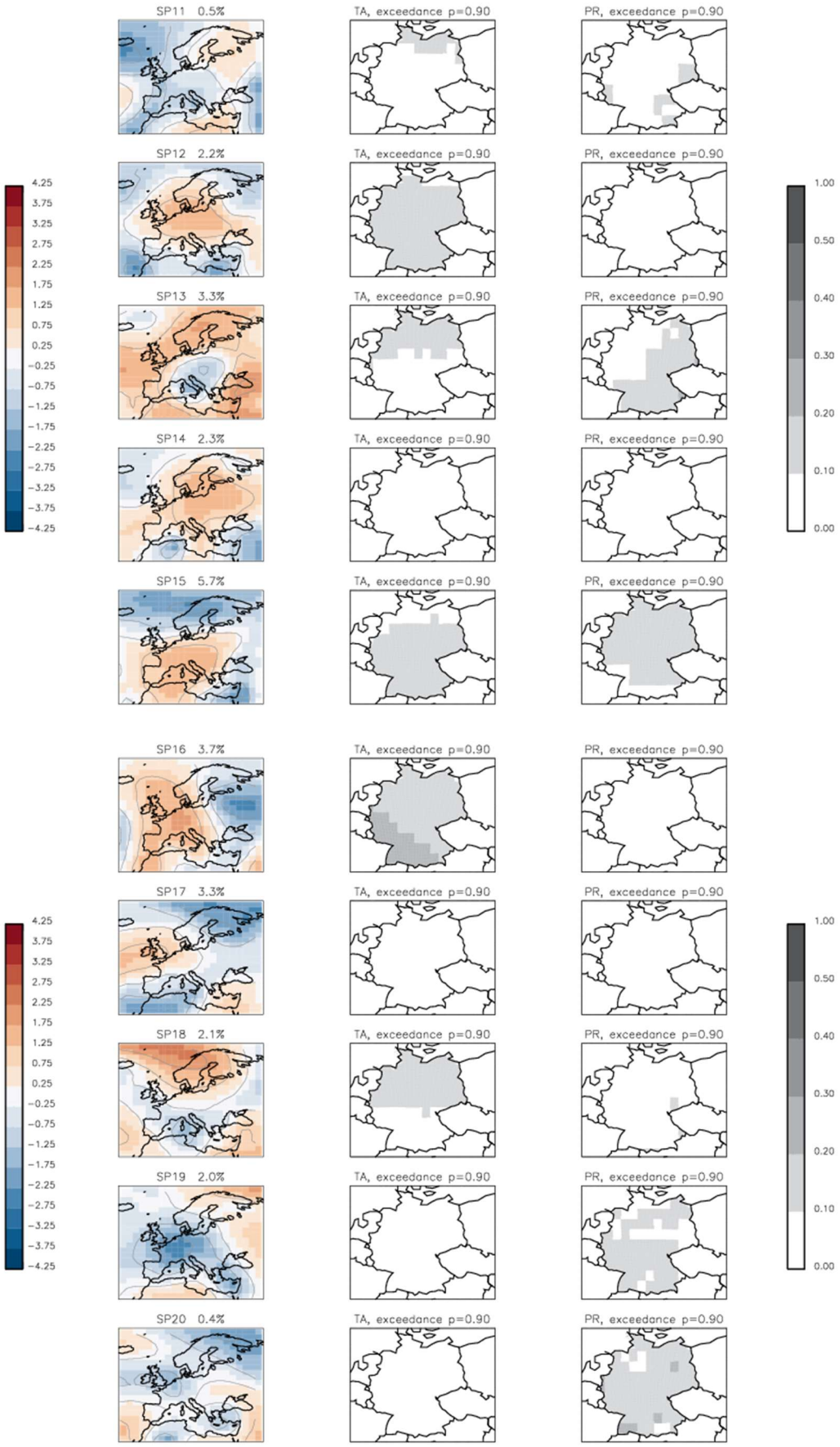
One of the major motivations given for the use of a large number of clusters in full field data was that this would help capture extreme events, whereas PCA based approaches with smaller numbers of clusters may not capture extremes. Unfortunately, neither side of this claim has been demonstrated. I also have some reasons to doubt the claim: looking at their figure 8, there is not much sign that the more common weather patterns are less extreme supporting than the rare patterns. Further, other work has shown that persistent regimes (i.e. common weather types) can drive cold and warm extremes.
As I suggested previously:
"I suggest that the authors more tightly focus the structure of the article around the importance of handling rare synoptic conditions and extremes in clustering approaches, showing an example situation where an impactful event was linked to a very rarely occurring circulation as motivation. I would then suggest a concrete demonstration that the EOF Kmeans with MSE approach more poorly handles rare circulations than the SSIM approach in ERA Interim…."

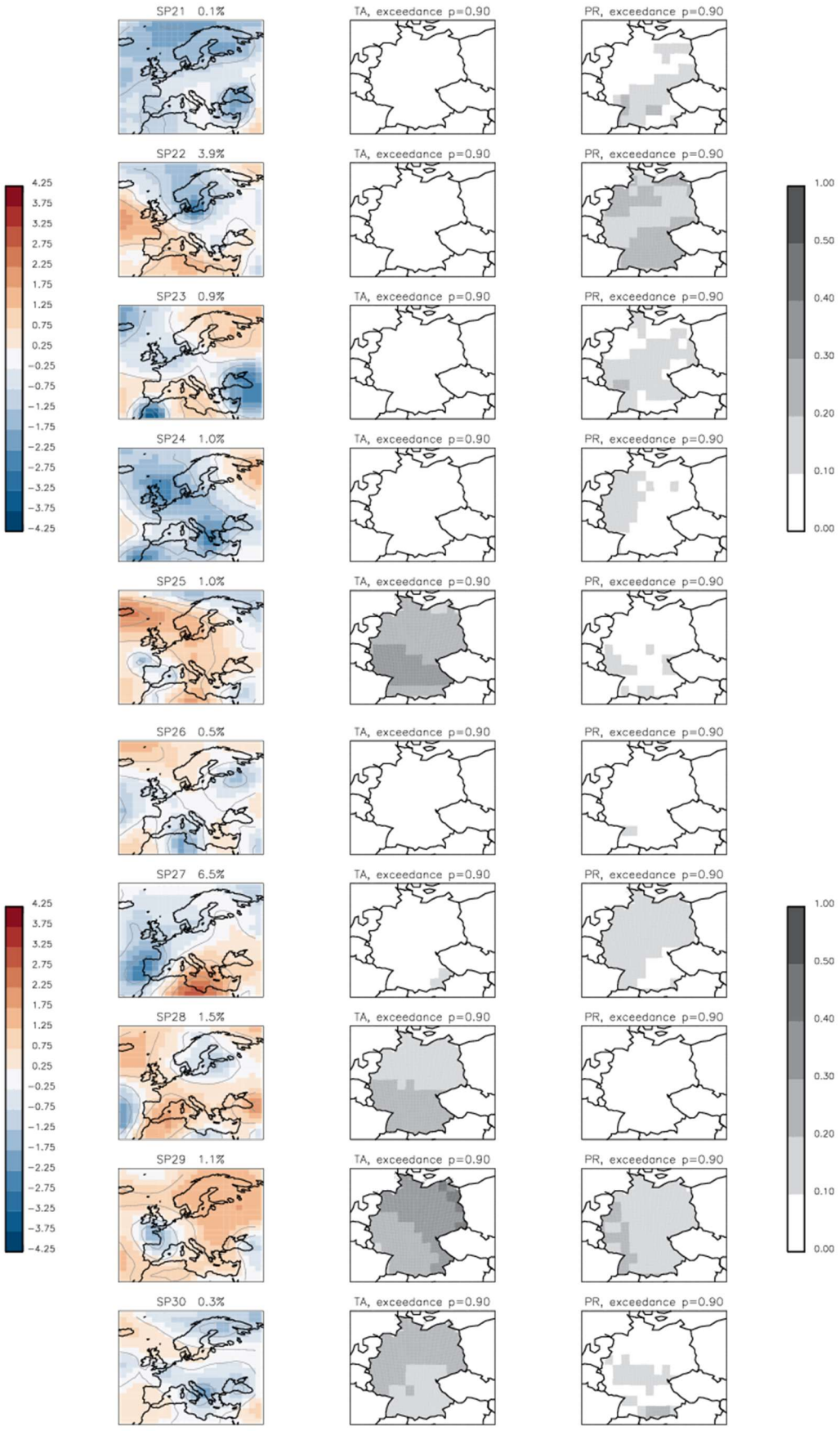
We would like to demonstrate on the example of Germany how having many classes of synoptic patterns may help to capture extreme events. Apart from the type of extreme events mentioned by Reviewer 1, which materialize through persistence of possibly not very rare circulation types, there are others that are related to rare circulation patterns. In the figure below we show each synoptic class (left plot), the fraction (of total elements in the class) that exceed the 90-percentile near surface temperature (middle plot) and the fraction that exceed the 90-percentile in total precipitation (right plot). All data – temperature and precipitation percentiles - we computed on the ERA-Interim 1979-2018 data as the zg500 used for the classification. We validated these percentile values with the corresponding percentiles of the Germanys national HYdrological RASterdata (HYRAS) data set (https://www.dwd.de/DE/leistungen/hyras/hyras.html) and found them to match regularly (we do not show the HYRAS percentile exceedances here now, but are ready to show them upon request).

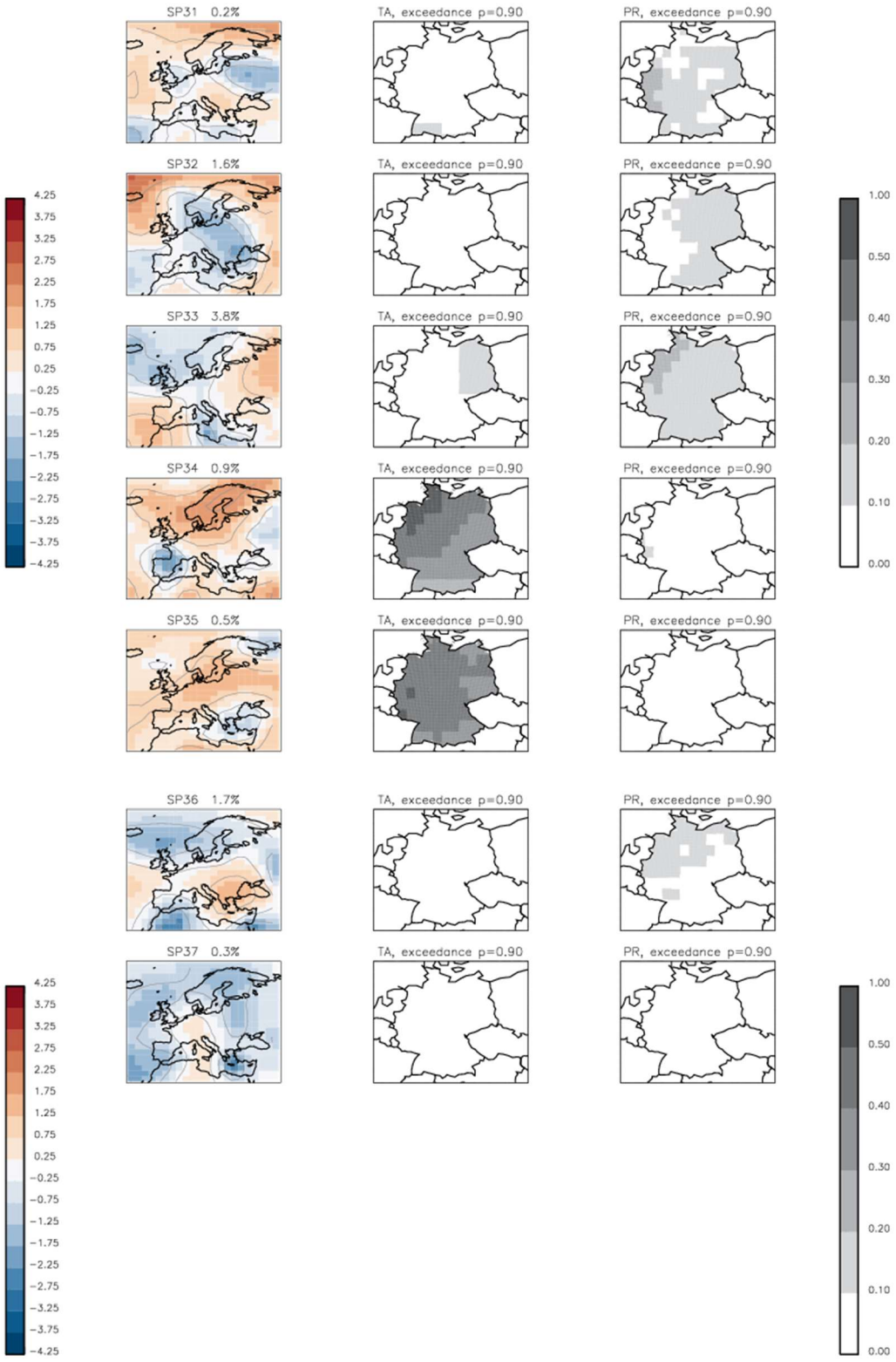
We show plots for all classes as we think it is necessary to demonstrate that some classes have no "extreme" events (SP3, 5, 9, 14, 17, 37), but some others do. Rare classes (with occurrence of less than 2% in total data) SP 25, 28, 29, 30, 34 and 35 are often "hot" i.e. show exceedances of 90-percentile in temperature. Rare classes SP 19, 20, 21, 23, 24, 29, 30, 31, 32, 36 show exceedances of 90-percentile in precipitation.

Precipitation is especially "difficult" variable to evaluate in models. Dry/wet biases in models may result from bad physical parameterisations or/and from models disability to reproduce the correct synoptic pattern. Therefore, knowing that a particular synoptic pattern often goes along with strong precipitation, we can check if a model is able to reproduce this pattern or not. This knowledge would help to attribute precipitation errors to errors in models physics or dynamics. Having only fewer synoptic patterns would hamper the differentiation between the classes and not allow us to make such attribution.


SP1 11.6%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP2 4.4%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP3 4.9%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP4 5.2%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP5 3.7%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP6 8.0%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP7 3.7%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP8 1.8%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP9 3.1%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP10 2.4%
TA, exceedance p=0.90
PR, exceedance p=0.90
4.25
3.75
3.25
2.75
2.25
1.75
1.25
0.75
0.25
−0.25
−0.75
−1.25
−1.75
−2.25
−2.75
−3.25
−3.75
−4.25
1.00
0.50
0.40
0.30
0.20
0.10
0.00


SP11 0.5%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP12 2.2%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP13 3.3%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP14 2.3%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP15 5.7%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP16 3.7%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP17 3.3%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP18 2.1%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP19 2.0%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP20 0.4%
TA, exceedance p=0.90
PR, exceedance p=0.90
4.25
3.75
3.25
2.75
2.25
1.75
1.25
0.75
0.25
−0.25
−0.75
−1.25
−1.75
−2.25
−2.75
−3.25
−3.75
−4.25
1.00
0.50
0.40
0.30
0.20
0.10
0.00

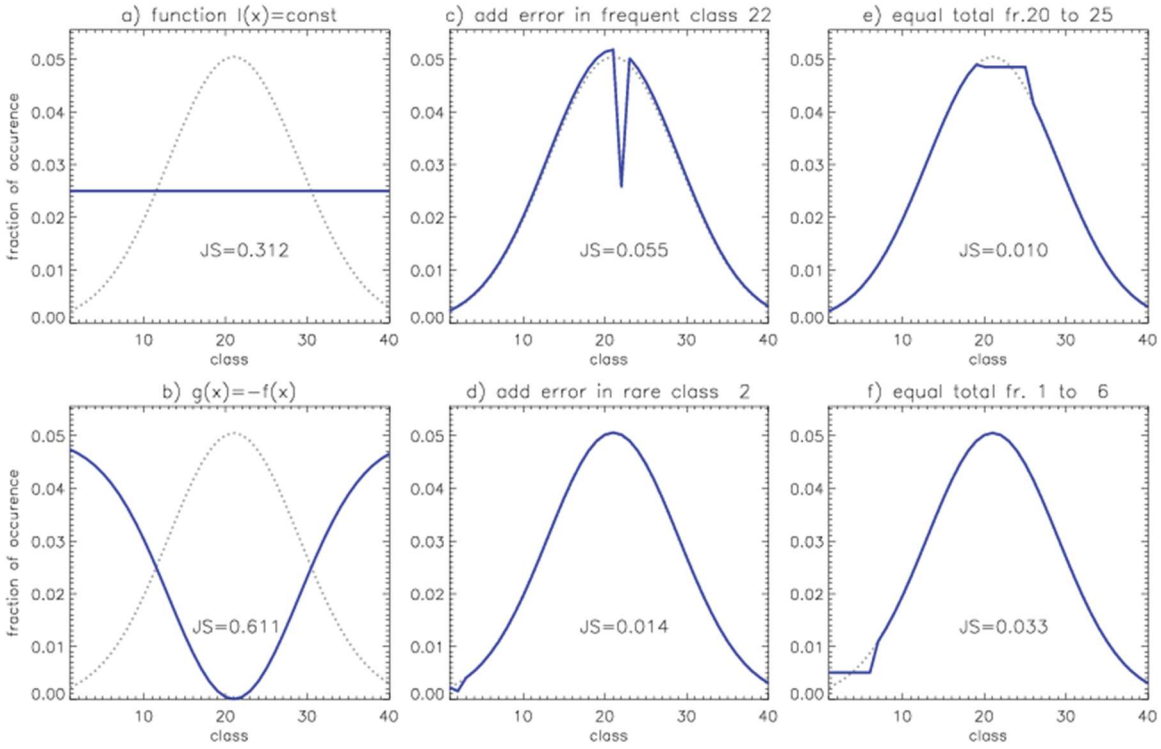

SP21 0.1%
SP22 3.9%
SP23 0.9%
SP24 1.0%
SP25 1.0%
SP26 0.5%
SP27 6.5%
SP28 1.5%
SP29 1.1%
SP30 0.3%
TA, exceedance p=0.90
PR, exceedance p=0.90
4.25
3.75
3.25
2.75
2.25
1.75
1.25
0.75
0.25
−0.25
−0.75
−1.25
−1.75
−2.25
−2.75
−3.25
−3.75
−4.25
1.00
0.50
0.40
0.30
0.20
0.10
0.00


SP31 0.2%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP32 1.6%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP33 3.8%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP34 0.9%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP35 0.5%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP36 1.7%
TA, exceedance p=0.90
PR, exceedance p=0.90
SP37 0.3%
TA, exceedance p=0.90
PR, exceedance p=0.90
4.25
3.75
3.25
2.75
2.25
1.75
1.25
0.75
0.25
−0.25
−0.75
−1.25
−1.75
−2.25
−2.75
−3.25
−3.75
−4.25
1.00
0.50
0.40
0.30
0.20
0.10
0.00

Even if it is the case that rare circulations are associated with rare extremes, when you compute the Jensen-Shannon divergence, you weight each class by frequency!
So representation of rare flows has almost no impact on the resulting quality index.

We show examples of Jensen–Shannon distance values between a 40-element Gaussian-shape histogram f(x) as reference (shown in plots by the grey dashed line) and six alternative histograms:

a) histogram of equally frequent classes (uniform distribution l(x)=const)
b) "mirrored" histogram (g(x) = – f(x)+a)
c) histogram with reduced frequency of one frequent element
d) histogram with reduced frequency of one rare element
e) histogram with 6 equally distributed frequent classes (total frequency of these 6 classes kept constant)
f) histogram with 6 equally distributed rare classes (total frequency of these 6 classes kept constant)



From the above plot, we see that JS=0.312 between a Gauss-shape and a uniform distributions should be considered a very large distance value as the compared distributions are obviously very different. Please note, the Jensen–Shannon divergence is bounded by 1 for two histograms (using base 2 logarithm), and therefore, JS-distance is bounded by 1 as well.
Reviewer 1 argues that rare classes have no/small impact on JS-distance, referring probably to a situation as in the plot "d": an error in frequency of one rare class makes moderate contribution to the total JS-distance (as compared to the contribution of the similar error in one frequent class, plot "c"). The plots "c" and "d" may lead the observer to a conclusion that errors in rare classes are negligible in computing JS-distance. This is a misleading

conclusion. Please see plots “e” and “f” for clarification: JS-distance is higher in response to errors in multiple rare classes (plot “f”) as to the errors in frequent classes (plot “e”). The relative change in frequency of the rare classes by such error is quite high (the mean distribution M used in computation of Jensen-Shannon Divergence undergoes large changes relatively to the original distributions). The situation in our evaluation framework is most similar to plots “e” and “f”, where the JS-distance appears to work quite satisfactorily.

As the *Quality Indices* (computed on JS-distance) may look much of the same magnitude, we list here our comparison of models in terms of JS-distance:

Nr	*Model name*	*JS for individual statistics*							*Mean JS*
		HIST	*$HIST_{DFJ}$*	*$HIST_{MAM}$*	*HIS_{JJA}*	*$HIST_{SON}$*	*TRANSIT*	*PERSIST*	
-	ERAINT(ref.reanalysis)	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
-	NCEP (alt.reanalysis)	0,013	0,017	0,020	0,028	0,021	0,079	0,062	0,034
1	ACCESS-CM2	0,057	0,115	0,065	0,125	0,080	0,165	0,128	0,105
2	AWI-ESM-1-1-LR	0,072	0,097	0,092	0,126	0,114	0,170	0,125	0,114
3	BCC-CSM2-MR	0,061	0,096	0,085	0,140	0,111	0,168	0,122	0,112
4	BCC-ESM1	0,067	0,113	0,106	0,143	0,104	0,171	0,124	0,118
5	CanESM5	0,061	0,124	0,097	0,091	0,096	0,174	0,128	0,110
6	CESM2	0,064	0,093	0,081	0,116	0,101	0,164	0,126	0,107
7	CESM2-FV2	0,079	0,125	0,087	0,138	0,120	0,181	0,136	0,124
8	CESM2-WACCM-FV2	0,074	0,118	0,113	0,151	0,089	0,174	0,132	0,122
9	CMCC-CM2-SR5	0,073	0,111	0,080	0,161	0,100	0,176	0,125	0,118
10	CNRM-CM6-1	0,059	0,105	0,081	0,150	0,088	0,169	0,128	0,111
11	CNRM-ESM2-1	0,043	0,098	0,087	0,119	0,089	0,164	0,126	0,104
12	EC-Earth3	0,054	0,091	0,076	0,137	0,095	0,164	0,120	0,105
13	EC-Earth3-Veg	0,068	0,091	0,081	0,165	0,085	0,170	0,117	0,111
14	FGOALS-f3-L	0,068	0,147	0,104	0,173	0,076	0,170	0,124	0,123
15	FGOALS-g3	0,073	0,141	0,097	0,145	0,081	0,175	0,138	0,121
16	GISS-E2-1-G	0,061	0,127	0,097	0,178	0,093	0,171	0,120	0,121
17	HadGEM3-GC31-LL	0,050	0,108	0,078	0,107	0,086	0,161	0,132	0,103
18	HadGEM3-GC31-MM	0,054	0,090	0,084	0,116	0,077	0,163	0,122	0,101
19	INM-CM4-8	0,071	0,106	0,096	0,170	0,110	0,182	0,136	0,124
20	INM-CM5-0	0,059	0,089	0,095	0,121	0,123	0,166	0,139	0,113
21	IPSL-CM6A-LR	0,065	0,099	0,099	0,181	0,131	0,169	0,132	0,125
22	IPSL-CM6A-LR-INCA	0,056	0,124	0,094	0,176	0,131	0,168	0,136	0,126
23	KACE-1-0-G	0,051	0,090	0,081	0,125	0,079	0,163	0,130	0,103
24	MIROC6	0,063	0,105	0,076	0,136	0,094	0,164	0,136	0,111
25	MPI-ESM-1-2-HAM	0,061	0,104	0,085	0,127	0,104	0,168	0,122	0,110
26	MPI-ESM1-2-HR	0,057	0,105	0,082	0,098	0,088	0,166	0,118	0,102
27	MPI-ESM1-2-LR	0,056	0,103	0,070	0,112	0,085	0,164	0,124	0,102
28	MRI-ESM2-0	0,052	0,090	0,098	0,122	0,079	0,161	0,118	0,103
29	NorESM2-LM	0,077	0,124	0,134	0,175	0,126	0,180	0,142	0,137
30	NorESM2-MM	0,065	0,108	0,087	0,127	0,126	0,172	0,129	0,116
31	TaiESM1	0,060	0,121	0,091	0,119	0,091	0,166	0,134	0,112
32	UKESM1-0-LL	0,060	0,073	0,082	0,139	0,089	0,161	0,128	0,105
-	*MEAN (32 models)*	0,062	0,107	0,089	0,138	0,098	0,169	0,128	0,113
-	*STDDEV(32 models)*	0,008	0,016	0,013	0,024	0,017	0,006	0,007	0,009

The frequency histograms of the projections show JS distances around 0.1 (TRANSIT and PERSIST around 0.15), which is of course not way off the reference but noticeable. It would be interesting to compare that to earlier model generations to appreciate the improvements.
The computation of the *Quality Index* based on the JS-distance was done for convenience for our downstream application, which admittedly has no relevance for the contents of this manuscript:

$$QI(P \parallel Q) = exp^{-a\sqrt{JS}}$$

Besides, *QI* can be scaled by the constant *a* for a better differentiation among models. If Reviewer 1 considers it appropriate, we can skip the transformation completely.

Usefulness for climate model evaluation

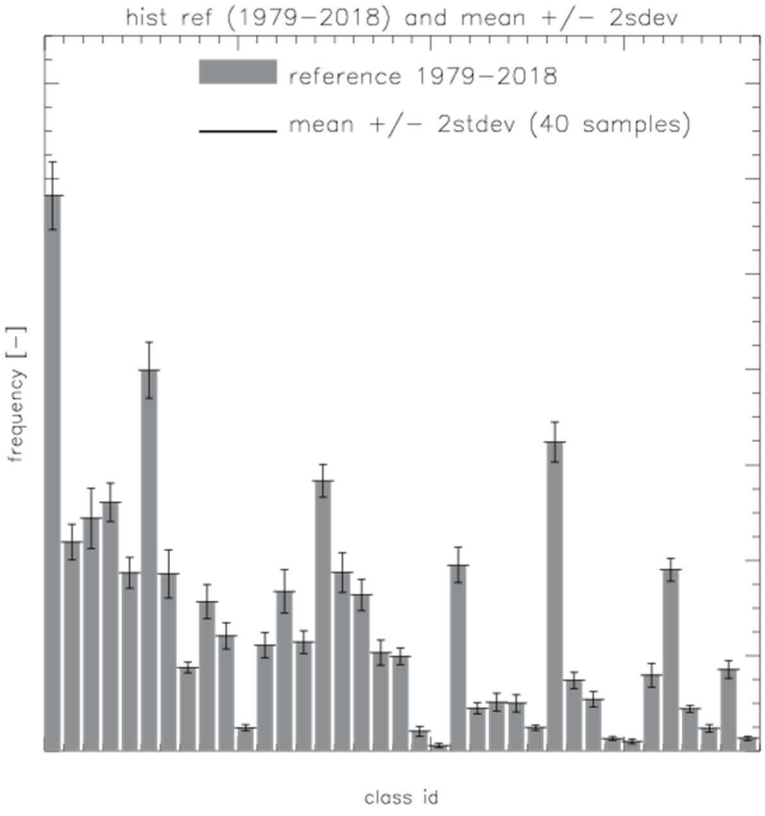
The authors also emphasise the value of their method for climate model evaluation. Indeed, circulation based metrics can be very useful for such analysis. This can and has been done several different ways (although it would be easy to think otherwise reading the authors' work), with only a few regimes at one extreme as in [1], or on a gridpoint basis as in [2] at the other extreme.
However, I am seriously concerned that the method the authors suggest is not suitable for this purpose.
The author's explain that using similarity as a metric, ~37 weather patterns are needed to fully capture the diversity of European circulations. I accept this, and it is a useful perspective, and similarity is a nice way to quantify this. Exploring spatial and seasonal variations in this number of 'necessary patterns' could be an interesting dynamical study. But, for model evaluation, the question of relevance is not how many weather patterns you need, **but how many weather types you can constrain**, given data limitations.

It is generally acknowledged that MSE/Euclidian distance is not appropriate for use in high-dimensional spaces. Therefore, the use of prior dimension reduction is usually recommended. This drawback is less pronounced for SSIM, which makes it an alternative measure, when dimension reduction is undesirable.
All 37 classes presented in Figure 8 may look "patchy" and not different enough from each other at the first glance. However, all these classes are not similar (according to our definition) to each other as each pair of them has similarity value smaller than 0.40 (the threshold chosen for the 2-step classification algorithm). Two points are important to note: 1) This class separation is done in terms of SSIM and does not have to hold in terms of MSE. It might occur that MSE cannot differentiate between two patterns, but SSIM can. 2) Each class is represented by its medoid - this makes the separation of classes sharper and the assignment of samples less ambiguous as compared to the common practice of using centroids. The attribution of each data element to a class is done using SSIM with respect to the medoids.

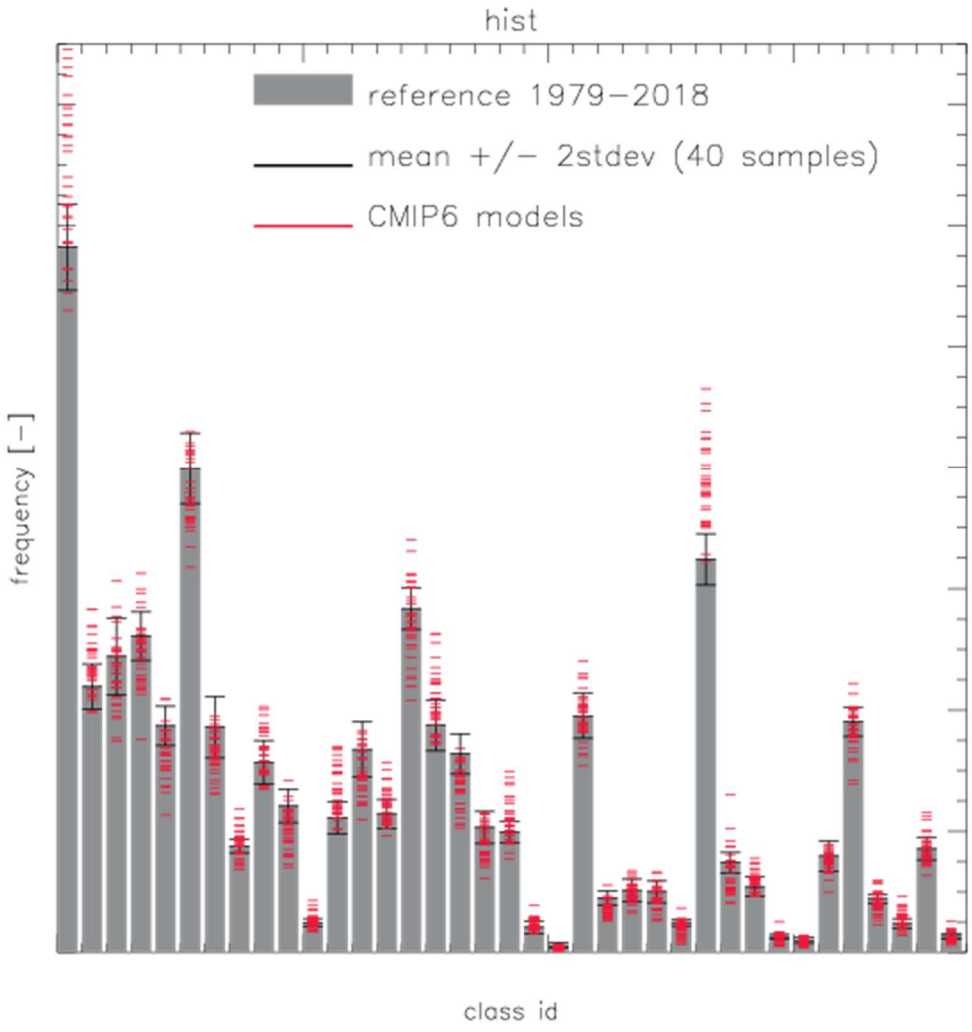
The authors compute error metrics for weather pattern frequency (37 elements), transition matrix (37x37 =1369 elements) and persistence probability over days 1-8 (37x8=296 elements). Simply put, using 40 years of ERA-Interim the sampling uncertainty in such fine-grained metrics are almost certainly far larger than any difference between climate models and era-interim. The fact that the inter-model variation in scores is so low reinforces this

point. I believe your quality index is almost entirely noise, averaged over a few hundred variables.
I make this claim quite confidently, as I know that it is difficult to find significant differences in the frequency and persistence of models and reanalysis when only using 3-10 regimes, and 100 years of data. Of course I would be pleased to be proven wrong: if you can rigorously constrain sampling variability in model and observational statistics, and so provide upper and lower bounds on your quality index, and still get meaningful results, then the scientific contribution is strong. Otherwise, I would move away from climate model evaluation as a goal for this methodology.

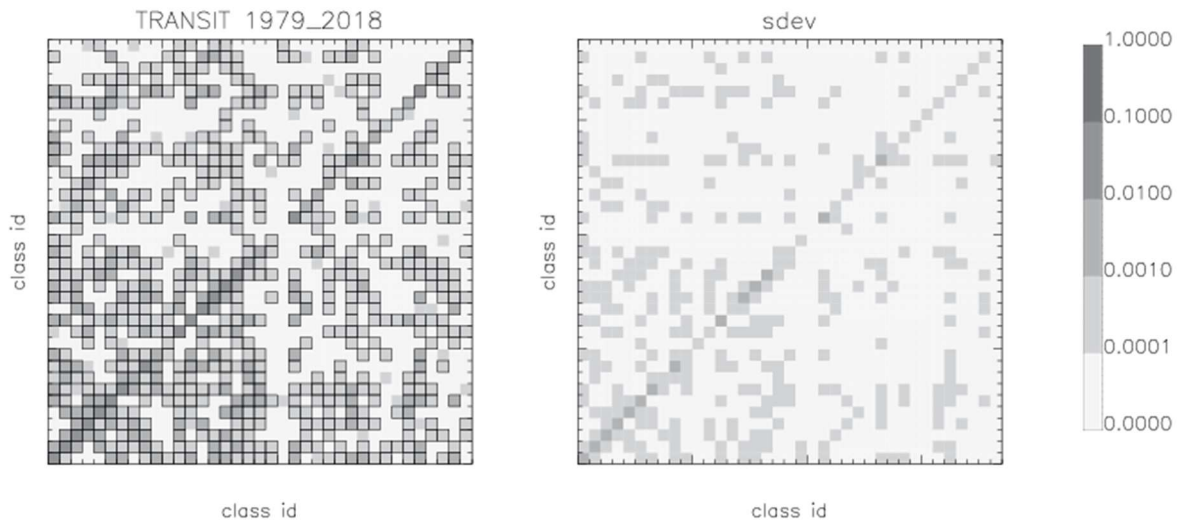
As it is conceptually difficult to assess sampling uncertainty with only one realisation (which is furthermore not a simple Multinomial distribution), we use resampled data (10-fold block-cross validation, i.e. a sliding block of 10 years cut from the sample in a cyclic way) as to build 40 different sets of 30 year-histograms HIST, HIST_JFD, HIST_MAM, HIST_JJA, HIST_SON, and TRANSIT, PERSIST-matrices. From these 40 histograms we estimate standard deviations for each element (frequency/persistence of a certain class, transition probability from one class to another) of these one- and two-dimensional histograms.
As a very rough, zeroth-order check of robustness we compare the estimated values in the frequency histograms and the TRANSIT/PERSIST matrices with two times their resampling standard deviation. It appears that the uncertainty in the histograms is reasonably low: all values in the histogram are greater than their individual 2*std, even for the rare classes. This is in line with our earlier observation that the clustering algorithm stabilizes at around 20 years of daily data.



Furthermore, the values produced by the CMIP6 projections, though some do fall inside our "confidence intervals", show far higher departures from the reference than the resampled data, both in frequent and rare classes. Interestingly, many models seem to shift weight from a number of rare classes to the frequent ones, which would indicate a reduced diversity of circulation types (see figure below). In our judgment, this histogram corroborates the usefulness of our approach comparing the frequency histograms of the projections to the reference.

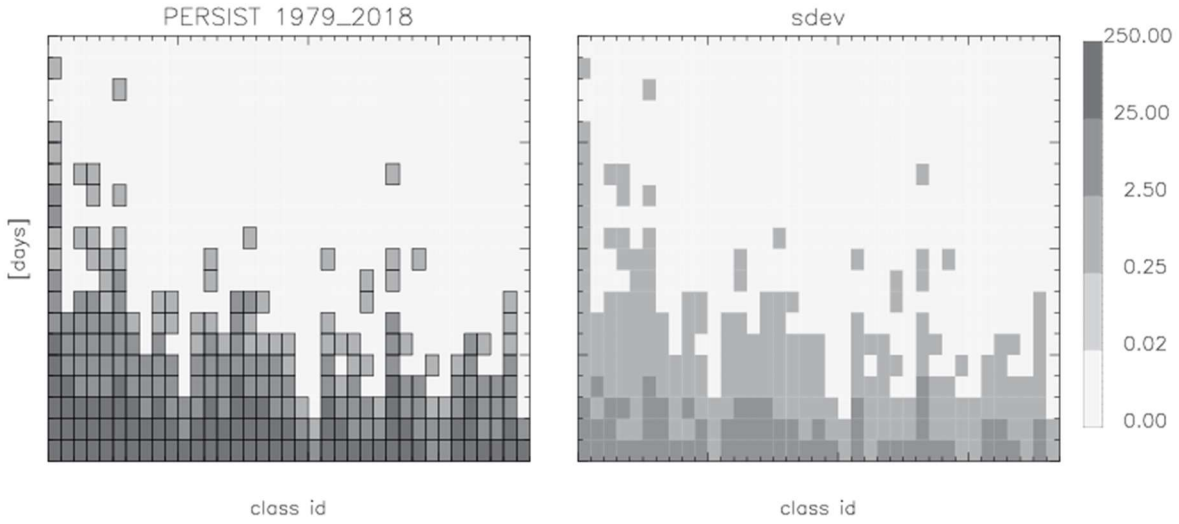


TRANSIT and PERSIST matrices.
As the Reviewer 1 suggested, some elements of the matrices TRANSIT and PERSIST do not satisfy our requirement of robustness. We show the TRANSIT-matrix along with its resampled standard deviation in the following figure (matrix elements highlighted with black contour lines are >= 2*std, std is computed over 40 resamples of TRANSIT matrices). Note that the scale is logarithmic here.



From the above figure we see that transition-elements with larger absolute value (higher frequencies of occurrence) are likely to be robustly estimated, small elements (rare transitions) are less likely so, as expected.

In contrast, the PERSIT-matrix seems to be sufficiently robust, i.e. all elements are greater than their 2*std. This might result from the lower number of elements contained in the matrix.



We suggest, that a larger data set will certainly help us to estimate all elements of the TRANSIT-matrices more robustly. For now, we agree that using about 14600 data elements

may be not enough to robustly estimate the sampling uncertainty for all elements in the transition matrix. Nonetheless, TRANSIT is certainly not “all noise”.

Synthetic data

The synthetic data section raises some questions for me. One clear point that I found interesting is that K-means leads to distorted patterns (i.e. not circles as in the synthetic data). However, I think the other points would be better made in ERA Interim than in the synthetic data. The synthetic data does not have multimodal structure, so there is no reason to expect any clustering algorithm to give very clear clusters: there are no clusters to identify, just ‘hallucinations’ of the method. In fact, you could argue that in non-structured data, a good clustering algorithm *should* give unclear structures.

It is true that the synthetic data are generated randomly and have no genuine cluster structure of zg500. But all clustering algorithms that we know of, would produce clusters governed by the position of the largest anomaly in the domain and its sign. In the synthetic data example: the initializing clusters are produced by the HAC, which has no predefined number of classes but is only driven by the similarity threshold. Even in a completely random dataset clusters can be constructed because some samples are more similar to each other than others. In this respect, our clustering algorithm is neither better nor worse than any other algorithm.
We have chosen to show the performance of the clustering algorithms (with k-means/k-medoids using MSE/SSIM) for demonstrating the distortions of cluster centres that occur when using means, but cannot by construction occur with medoids. Using real data, these distortions would be less obvious.

Also, I do not follow the claim about snowballing: the k-medoids with SSIM produces the most snowballing of all algorithms shown in figure 4.

We don’t see why Reviewer 1 is claiming that k-medoids with SSIM is producing the most “snowballing”. The “snowballing” does not mean forming classes with many elements but classes with “vanishing structure”, which is not the case in Figure 4 d. We explained this in the manuscript lines 317-325:
We already showed (Figure 1 and Figure 3) that small MSE does not guarantee the structural similarity of compared patterns. Classes built with k-means-MSE show very little structural detail as a result of building cluster centroids over multiple class elements, whose structural similarity remained unaccounted. The danger of having such classes “with vanishing structure” is that they may serve as attractors for further elements as the clustering algorithm runs targeting at minimizing MSE only. This leads to the so-called “snowballing” effect i.e. the more elements are assigned to this class, the less structure shows its centroid, the more elements are assigned and so on. Cluster 9 (Figure 5) is a good example of such “snowball”-class: although all shown elements have comparable small MSE to the final class centre, their visual (for an 13 observer) and computed similarity (value of SSIM) differs strongly as shown for a group of the first 28 elements (out of 132) indicating a strong structural inhomogeneity of patterns contained in one class. This example demonstrates the danger of building “snowball” classes when using MSE as distance metric for data with highly structured patterns.

In the classification with the ERA-Interim data, the Figure 14 shows the high similarity between the class medoids and their centroids (mean of all class elements). This means that these classes are not “snowballs”. Although the classes may have many members, they show pronounced and similar (within the class) structural patterns.