

Responses to Reviewers for WCD paper

Response to Reviewer #1:

The manuscript presents a novel atmospheric blocking detection method based on Self Organizing Maps (SOM), but most importantly compares it against a novel ground truth for European summer blocking which has been defined in both model and observations. Especially this latter point is very interesting since it provides a quite “revolutionary” and unique dataset to work with. The manuscript is interesting and presents a novel approach to the widely known but long-lasting issue of objective blocking detection. The SOM method seems to provide better results than the currently used detection method, although some unclear passages in the presentation makes unclear to me to what extent the improvements presented are actually useful in climate model assessment. The presentation is overall clear, even in some instances some sentences are hard to follow and might need some rephrasing.

We thank the referee for the supportive comments and helpful critique of the manuscript. In response to this comment, we have also revised a few paragraphs to be more concise and clearer in their wording.

As also commented by the author, the absence of a gridded dataset as an output is a quite significant limitation, since it implies that the SOM blocking index is valid only a certain box – and the definition of the threshold such box is arbitrary. Therefore, I wonder how can be defined - using the SOM-BI - the blocking impact over a certain European subdomain? Or how can we extract information on the Rossby wave breaking properties associated with a specific blocked event? These are only a few examples of the usages of the traditional blocking indices, so I am wondering to what extent this interesting and innovative approach can be used to investigate climate models or atmospheric blocking impact. It would be nice if the authors can comment more about this.

Thank you for this comment, because it is indeed a good idea to illustrate in our manuscript how the SOM-BI could be used to study blocking over, e.g., a certain European subdomain. The SOM-BI identifies blocked events using groups of SOM nodes. These contain spatial information about the location of the circulation anomaly. To illustrate how the SOM-BI information can be used for spatial analyses, we have

- (a) added a new section to the manuscript to describe how the SOM-BI can provide this information (section 2.7), and an illustrative application of the SOM-BI to study regional trends in blocking using ERA5 data has been included in section 3.7.
- (b) added a gridded blocking climatology to the manuscript that combines the improved detection skill of the SOM-BI with the DG83 index. Such a spatial SOM-BI climatology is based, only on those days where the SOM-BI identifies a blocking pattern. This climatology is shown in the bottom panel of Figure 12c. Similarly, other indices or dynamical metrics could be combined with the improved detection skill of the SOM-BI.
- (c) extended the SOM-BI in a new way to study European subdomains, by classifying different types of European blocking events. Since each blocked node group from the SOM-BI is associated with a unique spatial pattern (Figure 5), these node

groups can be grouped/post-processed using K-means clustering analysis to identify a set of SOM node clusters that are associated with blocking patterns. Figure 12 shows the case of 4 clusters, which highlight 4 distinct spatial patterns of European blocking (N, NW, W, E). As shown in Figure 12 (d), these can be studied independently to provide detailed information on the blocking impact over a European subdomain. We note that this post-processing provides entirely different patterns than if K-means would be applied to the raw data (Fig. A4 (a)). To explain this, we further discuss the differences between the application of K-means and SOMs to the raw climate dataset below, and we highlight that K-means is often applied to SOM nodes as a post-processing step (e.g. Cabanes and Bennani (2008), Deetz (2019)).

As already indicated in point (b), information on the Rossby wave breaking properties cannot be directly derived from the SOM-BI, but methods that are designed for studying these features can be integrated into the SOM-BI to show the Rossby wave breaking features of different types of blocking events. A discussion of the utility and further application of the SOM-BI is included in the discussion section (L 586-600):

"The use of SOMs as a blocking index provides opportunities for regional study that are not directly available in the other BIs. Through an additional post-processing step involving K-means clustering on blocked node groups (sections 2.7 and 3.7), we have shown that the SOM-BI can identify specific types of blocking events and provide detailed information about the changing nature of blocked events over a European subdomain (Fig. 12). The case of $k = 4$ has been shown in Fig. 12, but larger values of k can also be chosen to identify more distinct types of blocking pattern. Whilst the SOM-BI does not directly produce a gridded climatology of blocking patterns, we have shown that the SOM-BI can be integrated with the other BIs to develop a climatology that only considers only those days detected as blocking by the SOM-BI. This results in a SOM-BI climatology with a higher precision than the BI climatology."

We intend to apply this method to future trends across CMIP5 and CMIP6 models to better understand the patterns of blocking in models, diagnose model skill at reproducing the historic patterns of European circulation regimes and compare projections of future changes in blocking patterns. The identification of distinct blocking patterns from node groups enables a detailed study of blocking characteristics over European subdomains as shown in Fig. 12. Further quantities such as the Rossby wave breaking properties or the nature of blocking onset and decay can also be studied. This could be done by studying particular dynamical quantities on the blocked days identified by the GTD or the SOM-BI, and extended by contrasting the dynamical quantities across different categories of blocking pattern identified by the SOM-BI node groups."

Finally, we note that we do not consider the definition of the domain to be arbitrary. It is rather motivated by our specific interest in the European sector. We have added an explanation for our choice of domain in L 145-148: "JJA Europe was chosen because of our interest in the role of atmospheric dynamics in the development of mid-latitude land heat waves. Europe is a region which has seen many recent significant heat extremes (Christidis et al., 2014), and the role of changes in atmospheric dynamics in those heat extremes has been an active area of research (Cattiaux et al., 2013; Horton et al., 2015; Saffioti et al., 2017; Huguenin et al., 2020)."

In the same way, it is unclear to me how the comparison with the traditional blocking indices has been carried out: did the authors consider the whole European sector as blocked if only one grid point is blocked? The description of this method is

presented in an overly simplified way at L330. This approach may explain the incredibly high frequency of blocking since for AGP (92%), which in reality is about 10 times less over Central Europe, as also shown in Figure A3 by the authors. This passage should be clarified.

We define a day as blocked in the ground truth dataset if there is a block somewhere over the considered European domain. All blocking indices have only been calculated within the European domain, so all thresholds for amplitude, area, persistence and overlap need to be met within the European sector. This eliminates any events on the edge of the domain that we are not concerned with in this study. The discussion in section 3.2 has been expanded and clarified across L381-386:

"For our comparison, we first apply all BIs to the historical ERA5 data over the European domain. Each day for each BI is labelled as blocked if a blocking event has been identified within the European sector and persists for at least five days. A blocking event is not identified if the thresholds for amplitude, persistence, area and overlap discussed in section 2.3 are not met within the European domain. This results in a binary dataset for each BI that identifies periods of at least five consecutive days where blocking patterns exist within the European sector. These binary BI data sets we then compare to our manually labelled GTD."

We have also revised the AGP calculation using the method presented by Woollings et al. (2018), discussed in more detail in response to the next comment below, which has addressed the higher frequency of blocking detected by the AGP index due to the influence of subtropical highs. Accordingly, Figure A3 and the numbers in Table 1 have both been updated. The key updates to Table 1 are:

- Frequency of blocking for the ERA5 AGP index has been revised from 92% to 59%.
- Precision for the ERA5 AGP index has increased substantially from 0.40 to 0.84.
- Recall for the ERA5 AGP index has decreased from 0.78 to 0.48.
- F1 score for the ERA5 AGP index has increased from 0.53 to 0.61.

The comparison with the application of the AGP index to mask blocked regions south of 40° N has been removed since it is now no longer necessary.

Note that whilst this correction led to significant improvements in the skill in the ERA5 case, the subtropical high was not observed in UKESM so the correction has not been applied. However, the numbers for UKESM AGP in Table 1 and the climatology in Figure A3 (d) have been corrected to fix an error in the longitudes of the domain. These changes to Table 1 for UKESM AGP are relatively minor with:

- an increase of precision from 0.27 to 0.29
- an increase recall from 0.29 to 0.41.
- an increase in F1 score from 0.28 to 0.34.

The AGP climatology for UKESM in Figure A3 (d) now shows blocking across central and Northern Europe and no subtropical high. Lines 202-205 have been added to the discussion of the blocking indices methods to reflect these changes:

“Finally, to remove the well-known problem of the AGP index identifying anomalous blocking events associated with the subtropical high in summer (Davini et al., 2012), we adopt the extra threshold of the AGP index from Woollings et al. (2018). The subtropical high feature was not observed in UKESM over Europe, since the zonal gradients have a smaller magnitude, so the standard AGP index is used for UKESM.”

It must be kept in mind that indices that the authors are using for comparison have been developed to study mainly winter blocking, and some of them are known for not being very suited for summer study. Did the authors perform the same analysis on winter blocking? One clear example is the AGP here used, which – as the authors noticed – produce a lot of noise at low latitudes. There are versions of this index which have been tuned to run also in this season. I would encourage the authors to make use of one of the improved versions – as the one presented by Woollings et al (2018) - which gets rid of the spurious blocking events at low latitude.

We have not performed the same analysis on winter blocking, because our primary project interest is in weather extremes during summer, but we note that all three indices have been used with these thresholds for all seasons (e.g., Pinheiro et al. (2019)). We have taken on the recommendation of using the Woollings et al. (2018) improvement to the AGP index to correct for the noise at low latitudes, see also our reply to the previous comment, which improved its F1-score performance slightly (from 0.53 to 0.61; Table 1). We agree that the application of the SOM-BI method to the winter season could be an interesting topic of future work.

Although I must admit I am not an expert of Self-Organizing Maps, I wonder to what extent a SOM approach is different from a k-means clustering with a predefined number of patterns. In this sense, I also wonder how a canonical k-means clustering with k=4, a widely adopted methodology to study the Euro-Atlantic mid-latitude variability which is based on Z500 anomalies, will rank among the detection methods here presented. Indeed, k-means aims at detecting Scandinavian Blocking specifically so I would imagine that this approach might have a high skill, comparable to the SOM (although also k-means has been developed for winter circulation). It would be very interesting to see a comparison between SOM and the Scandinavian blocking regime at least for reanalysis.

The key difference between Self-Organizing Maps (SOMs) and K-Means clustering is that SOMs use a neighbourhood function, which for each algorithm iteration enables neighbouring nodes to shift towards the best matching node (Figure 3). As a result, in each iteration not only the best matching node (equivalent to the closest centroid in k-means by e.g. Euclidean distance) is updated, but also – to a degree - neighbouring nodes on the map. Consequently, with increasing node number, SOMs represent a smooth topology across the nodes from left to right and top to bottom on the final map, reflecting a realistic continuum of weather patterns. However, for small centroid/node numbers, the difference between K-means and SOMs is small, because the SOMs will still differentiate the range of possible weather patterns, but within a few nodes, making a smooth transition across the map of nodes impossible. Consequently, the results for SOMs and K-means approach each other for small numbers of nodes (n)/centroids (k), as shown in the revised manuscript in Figure A4 (a) for the case of k and n equalling 4. This figure is also shown below. Therefore if K-means for $k=4$ was used as a blocking index in the same way as done for SOMs in our manuscript, it would give similar results to those shown in Figure 7 for $n=4$.

However, the difference between K-means and SOMs is significant for high node numbers. We have derived from Figure 7 that the optimum node number is 20 nodes for the SOM. With k=20, the K-means algorithm will lead to a set of distinct weather regimes, whereas the SOM will create a continuum of weather patterns. The SOM-BI associates node groups with blocking patterns, which is consistent with the behaviour of the SOM as it represents a smooth topology. The comparison for the case of k and n equalling 20 is shown in Figure A4(b).

In L 245-252 the comparison between SOMs and K-means has been discussed:

"This property of SOMs is also the distinguishing feature between SOMs and K-means clustering. In the case of K-means clustering, each node is updated at each iteration independently and no neighbourhood function is applied. K-means tries to maximize differences between the centroids such that it does not learn a topology. This difference between K-means and SOMs is minor for low node numbers, since the sharp differences in spatial patterns are imposed on the SOMs and the neighbourhood function has a limited effect. For larger node numbers, the SOM topology becomes smoother and the K-means centroids remain distinct rather than representing a continuum of states, whereas a continuum is a more realistic reflection of the actual atmosphere (Skific 2012). A comparison between SOMs and K-means analysis for 4 and 20 node/cluster numbers is shown in Fig. A4."

I understand the intent of the authors is to provide a comprehensive approach on the topic of the mid-latitude variability, but the discussion on the sinuosity seems out of context, especially considering that this is a hemispheric diagnostic and the study is strongly regional. I would suggest the authors remove it from the introduction and from the figures, and perhaps replace it with the k-means clustering.

This suggestion has been taken on, with the sinuosity discussion removed and K-means clustering for the case of k=4 has replaced sinuosity in Figures 5, 6, A1, and A2. In L 209-212 some of the literature that uses K-means clustering to study mid-latitude variability has been cited and briefly discussed:

"K-means clustering analysis (Diday and Simon, 1980) has also been extensively used to study the Euro-Atlantic midlatitude variability and identify weather regimes (Vautard, 1990; Michelangeli et al., 1995; Cassou, 2008; Ullmann et al., 2014; Strommen et al., 2019; Fabiano et al., 2021)."

The comparison in the case studies for K-means analysis has also replaced the sinuosity discussion in L 350-357:

"Figures 6d and 7d show a K-means clustering analysis for Z500 anomaly fields in the case of 4 centroids. As described in section 2.4, the case of K-means with 4 centroids produces a similar set of weather regimes to SOMs with 4 nodes. Consequently, the K-means analysis exhibits a similar behaviour to the SOMs discussed above but distinguishing between fewer weather regimes. One weather regime indicating Scandinavian blocking consistently represents the 2003 European heat wave across Fig. 6d, but the Westward shift of the high pressure centre from Scandinavia on 31 July to the UK on 8 August 2003 is not described by 4 centroids. For the 2019 heat wave in Fig. 7d, all four weather regimes are represented, and the blocked period is primarily associated

with a mixed weather regime. This shows that the 2019 case is also not described well by K-means clustering.”

As long as I understand the authors conclude that the best skill is obtained making use of the Z500 field. However, this seems to be somehow implied by the fact that Z500 is the field which the authors have used to define the ground truth. If the blocked days ground truth have been defined on SLP or on PV, would the SOM maps always show the Z500 as the best choice possible? I am not asking to redefine the ground truth using a different variable – this should be a monster work – but it would be nice if the authors could comment on this and support in a stronger way their conclusions.

Figure 1 shows the variables used in the definition of the ground truth dataset (GTD), where 1(a) shows Z500 contours, 1(b) shows Z500 anomaly and 1(c) shows IPV at 350K. All of these maps have been studied for each five day period across JJA 1979-2019 to identify a blocked region. Similarly, since PV variables are not available from the CMIP6 archive we used SLP instead in Figure 2(c). Therefore, a range of maps and variables was used to identify blocking patterns so that we are convinced that the conclusion that Z500 leads to the best performance in detecting blocking is not circular.

We think that the reason that the Z500 variable is most effective is because it exhibits the best signal-to-noise ratio. This is reflected in Figure 10, which shows that for a given node number the Z500 SOM-BI identifies blocking patterns with a smaller number of node groups than other variables.

We have added a more detailed explanation for the skill of Z500 in L 564-568:

“We have further compared the performance of SOM-BI for a range of variables in both ERA5 and UKESM that have been classically used to study blocking (Figs. 8 and 9). We find that the best skill is obtained when applying SOM-BI to the Z500 field because it exhibits the best signal-to-noise ratio in blocking identification. This is reflected in Fig. 10, which shows that for a given node number the Z500 SOM-BI identifies blocking patterns with a smaller number of node groups than other variables.”

In the abstract the authors claim that the algorithm has no arbitrary threshold and this is an advantage compared to the other objective blocking method (L10): although this is true in the strict definition of threshold, there are several arbitrary “decisions” that have been undertaken by the authors, as the domain definition, the number of modes used in the SOM, or the dataset on which to train the model (that although is shown to be weakly sensitive, it introduces a methodological difference). I don’t feel as any of these arbitrary decisions as feral issues, but I would encourage the authors to tone down their statement on the objectiveness of the algorithm since it is not radically different from the “standard” blocking indices they are comparing to.

The choice of domain is primarily guided by our study interest, and we use the standard IPCC definition of the domain (see our response to the first comment). The number of nodes used in the SOM is optimised (see Section 3.3 and Figure 8) to support the best performance. The choice of training dataset is also optimised (section 3.5 and Table 2). Therefore, we do not consider these decisions as arbitrary. Nevertheless, we have modified the wording concerning objective decisions in L 8-10:

"We then demonstrate that our method (SOM-BI) has several key advantages over previous BIs because it exploits all of the spatial information provided in the input data and reduces the dependence on arbitrary thresholds."

(“avoids the need for arbitrary thresholds” has been replaced with “reduces the dependence on arbitrary thresholds”)

Similar language has been modified in L 63-65:

"This SOM-BI method has advantages over previous BIs because it exploits all the spatial information provided in the input data and reduces the dependence on arbitrary thresholds"

The authors spent a lot of time describing the different sensitivities of the SOM-BI to the SOM parameters. This is certainly a good thing, but they do not focus on the dynamical meaning of the SOM-BI blocking index. It would be very interesting to see a composite of the geopotential height pattern (and/or on other dynamical fields) of the blocking events which compares the SOM-BI with other blocking detection methods, as well as the evolution on the onset and on the decay phase. I understand that the authors aim at having a robust index from the methodological point of view, but it is very important to see if the blocking identified shows physical characteristics which are reasonable.

Since we have developed a new method this paper has been primarily concerned with demonstrating that the SOM-BI algorithm is robust and performs well. The basic dynamical justification of the SOM-BI has been shown through the case studies, which show the labelling of the SOM-BI and the days blocked, alongside a set of SOMs. To develop the dynamical meaning of the SOM-BI, we have added sections 2.7 and 3.7 to extend the method and show the application of the SOM-BI, as described in our reply to the first comment. This has shown that specific dynamical information is present in the SOM-BI to provide regional information. The comparison with composite Z500 fields in Figure 12(b) to the composite fields derived from the node groups in Figure 12(a) shows that the blocking patterns identified by the SOM-BI show typical physical characteristics. This is now discussed in L 526-528:

"Fig. 12 (b) shows the mean Z500 field across all blocked days identified for each cluster, which are highly consistent with Fig. 12 (a). Consequently, the subsets of node groups are as expected physically consistent with the circulation patterns across these blocked days so that the K-means clusters can indeed be used to study specific types of regional blocking."

We agree with the reviewer that studying the dynamical meaning of the index further will be worth pursuing. We have in particular highlighted study of blocking onset and decay in lines 596-600:

"Further quantities such as the Rossby wave breaking properties or the nature of blocking onset and decay can also be studied. This could be done by studying particular dynamical quantities on the blocked days identified by the SOM-BI, and extended by contrasting the dynamical quantities across different categories of blocking pattern identified by the SOM-BI node groups."

Minor issues

L49: I found that a mention to weather regimes should be added in the introduction.

This has been added with reference to the K-means approach in L48-50:

“Other frequently used methods to study the climatology and characteristics of blocking include K-means analyses to study weather regimes (Vautard, 1990; Michelangeli et al., 1995; Cassou, 2008; Ullmann et al., 2014; Strommen et al., 2019; Fabiano et al., 2021)”

L49: Sinuosity also includes planetary waves oscillation which are not blocking, and it is a hemispheric diagnostic while blocking is regional. I would suggest removing the discussion on sinuosity here.

This discussion has been removed and replaced with the K-means clustering above.

L50: This sentence is a bit tangled up, please clarify.

This sentence has been revised in L51-53 to:

“It has been highlighted that consistency across various methods in detecting long-term changes is a necessary requirement to confidently identify trends (Barnes et al., 2014; Woollings et al., 2018).”

L53: typo “a SOM”

This typo has been corrected.

L54: please remove “and better instrumentation and observations”, it is out of context here.

This phrase has been removed.

L55: why are the authors referring to “surface” here?

This reference has been removed.

L58: what does the author mean here with “historical”? Please specify.

“Historical” here refers to the blocking patterns in the reanalysis period. The phrase “historical information” has been replaced with “reanalysis data” in L55-58:

“Here, we therefore define a new binary ground truth dataset (GTD) of European blocking events across June–July–August (JJA) 1979–2019, based on a five-day threshold, reanalysis data and expert judgement.”

L59: What the authors mean with sinuosity here? Are you referring to a specific measure of the “waviness” of the mid-latitude flow?

This reference to sinuosity has been removed consistent with above.

L64: typo, remove “of”

This has been removed.

L73: although is clearer later in the text, the concept of “blocking index skill” is quite new since to my knowledge a definition of ground truth for blocking events is quite uncommon. I guess that the authors need an introduction to the concept they refer to when they talk about skill.

This concept of blocking index skill has been briefly introduced in L 66-69 and a reference provided to section 2.6 where more detail is provided:

“We identify the skill of different BIs by developing a binary time series identification of European blocking patterns and comparing this to our GTD using standard skill metrics discussed in section 2.6. This study is the first to define a GTD and we use it as a benchmark to compare the skill of different BIs over a region.”

L84: this is more of a naïve question from my side: why do the authors refer to “hyperparameters” instead of simply “parameters”? It looks to me that the number of SOMs is a parameter of the blocking detection method, or I am missing something here?

The term “hyperparameters” is machine learning jargon to refer to the parameters that are tuned for the algorithm, which in this case includes the number of SOMs. In addition to the node number, there are other hyperparameters which have been optimized such as the algorithm initialization, learning rate and sensitivity of the neighbourhood function.

L98: the discussion of the anomalies should be left for the following part: here you are referring to which field you are using, so please remove the word “anomalies”.

The word “anomalies” has been removed here.

L108: So what is the cutoff frequency of the Fourier low pass filter?

73 days. This description in L130-132 has now been amended:

“This is a smoothed function of the 365-day seasonal cycle across Z500, VPV and Tsurf using the first six harmonics of their Fourier series, where the first harmonic corresponds to the mean and the fifth to a 73 day span.”

L111: The beginning of this sentence is clumsy: do the authors want to mean that a detrending is applied only over the region studied? It sounds a bit redundant information.

This sentence has been reworded in L134-136 to remove the redundant information:

“The Tsurf and Z500 anomaly fields in ERA5 have been detrended linearly across time to remove the effect of thermodynamic warming. Following Jézéquel et al. (2017) we subtract a spatially uniform trend, so that the horizontal gradients of the field are not altered.”

L121: I would remove “following IPCC AR5 definitions (Stocker et al., 2013). The northern latitude is extended to 76 N when using data on a 2x2 grid” and just saying that the region ends at 76N

We think that it is helpful for the reader to know that we simply follow IPCC AR5 guidelines here, see also our discussion concerning the choice of domain size above, so we have kept the text as is.

L122-124: this is one of the aspects of the authors' work I struggle to understand: as long as I see, the method works on pentads which are defined arbitrarily. What I do not understand is if there is or not a “running-mean” approach. As long as it is presented this method looks like a 5-day discretization of the dataset, for which e.g. day 1-5 are blocked, days 6-10 no, and so on. Blocking duration is thus a multiple of five? I downloaded the dataset from the Zenodo archive and it does not seem to be the case, so the authors will probably need to clarify this. I am not sure if I have misunderstood something here, but of course if a discrete approach has been chosen, this would be clearly a caveat because blocking is a continuous process, so that in this way you may lose some block events or consider some other which is only a partial event. As a consequence, the ground truth may be wrong and all the conclusions you are drawing may be re-discussed.

This appears to be a misunderstanding. Every five day period across JJA has been studied by eye and identified whether or not it is blocked across those five days. Then from these sets of classifications the GTD has been reconstructed, where a day is labelled as blocked if the previous four days have also been identified as blocked. So if days 1-5 are identified as blocked and days 2-6 are not, then days 1-5 will be labelled as blocked in the GTD and day 6 will be labelled as not blocked. This approach ensures that we consistently identify every blocked period without missing any events. We have added the following clarification to the discussion in L 159-163:

“Once the total set of all 4001 consecutive 5-day periods across JJA 160 1979-2019 has been classified, persistent blocking events are reconstructed to form a time series where each day is labelled as blocked or not. If a day belongs to any one of the consecutive blocked five day periods, it is individually labelled as blocked (1), and if a given day does not belong to any of the blocked five day periods it is labelled as not blocked (0). This creates a classification of blocking patterns for each day where each blocking event has a minimum length of five days.”

L380: why not use half of the dataset as training and the other half for evaluation? This is a more common approach I would say.

10-fold cross-validation, the approach we use here to split the data into training and test sets, is a standard statistical optimization method. Using half the dataset for training and half for evaluation would unnecessarily limit the number of training samples for the algorithm, and would also not allow us to evaluate the robustness of our algorithm for the entire historical reanalysis data (effectively splitting the data in half would be equivalent to one-sided two-fold cross-validation). We have also demonstrated (Figure 11) that the algorithm performs well when 10-20 years are used for training, so using half of the dataset as training and half for evaluation would result in a similar skill score. However, this would create fewer options to evaluate the robustness of our parameter settings across the reanalysis/model period. The shaded regions in Figure 11 show the standard deviation of the skill scores across multiple cross-validations on these 10 possible hold-out datasets.

Figure 5, 6: colors for the different blocking indices are missing in the caption.

Thank you, this has now been added in the caption for Figure 5.

Response to Reviewer #2

In this paper, the authors introduce a new blocking index based on self organizing maps. They compare this index to other commonly used blocking indices and to a ground truth time series they build based on expert judgement. They show that this new index generally performs better than other indices. I find this paper very well written and structured, with a very thorough sensitivity testing of their index to several key parameters. I have a few comments and questions that I list below. I recommend to accept this paper after some revisions.

We thank the referee for the supportive comments and helpful critique of the manuscript.

General comments:

You say your GTD is objective, but you make some heuristic choices too (namely the region and time period), so it's not completely objective. It's also a little bit unclear to me how you label your blocking periods from the maps shown in Figures 1 and 2. Do you manually look at each group of maps and identify blocking highs?

We actually do not consider the definition of the region to be an entirely arbitrary choice. The region we are studying is primarily motivated by our interest in the European sector. We have added an explanation for our choice of domain in L 145-148:

"JJA Europe was chosen because of our interest in the role of atmospheric dynamics in the development of mid-latitude land heat waves. Europe is a region which has seen many recent significant heat extremes (Christidis et al., 2014), and the role of changes in atmospheric dynamics in those heat extremes has been an active area of research (Cattiaux et al., 2013; Horton et al., 2015; Saffioti et al., 2017; Huguenin et al., 2020)."

The time period we have chosen is the longest time period currently available in ERA5. We have also shown in Figure 11 that the algorithm has reached an optimal performance when only 20 years are used, and after this point there is weak sensitivity to the algorithm's performance.

To label our GTD we manually look at each five day period such as those shown in Figures 1 and 2 and conclude by eye using all the variables shown if a blocking high persists across this period. This has been further clarified in L 149-150:

"The GTD has been derived by studying every successive five day period from 28 May to 4 September 1979-2019, and manually identifying whether or not a blocking high persisted across any such five-day period."

Then from these sets of classifications, the GTD has been reconstructed, where any day that has been identified as part of a blocking pattern in at least one of these five day periods has been labelled as blocked. This has been clarified in L 160-163:

“Once the total set of all 4001 consecutive five day periods across JJA 1979-2019 has been classified, persistent blocking events are reconstructed to form a time series where each day is labelled as blocked or not. If a day belongs to any one of the consecutive blocked five day periods, it is individually labelled as blocked (1), and if a given day does not belong to any of the blocked five day periods it is labelled as not blocked (0). This creates a classification of blocking patterns for each day where each blocking event has a minimum length of five days.”

Could you make it clearer in the text what’s the added value of the SOM-BI compared to the GTD? As it is presented, the GTD gives supposedly better results than the SOM-BI so why do you need to build the SOM-BI? This may be easier to understand once the methodology to construct the GTD is clearer.

The SOM-BI has been constructed from the GTD so that it can be used in climate models and future projections where no GTD exists. Since labelling the GTD is quite an onerous task, if an algorithm can be trained on the GTD, the labelling process can subsequently be automated, e.g. to apply the SOM-BI to subsequently collected new meteorological data or to extensive climate model data (as demonstrated here for the case of UKESM). The SOM-BI could further provide useful regional information when applied in such automated ways, because it classifies blocking patterns with different node groups which in turn have inherently different characteristics that reflect the stationarity and location of blocking patterns. We have also shown in Figure 11 that the SOM-BI can perform well when only 20 years are labelled, so future applications would not need the longer records (41 years and 101 years in reanalysis and model) to be labelled as done here. The longer labelling exercises demonstrate the sensitivity of SOM-BI to the length of training data, and the robustness of SOM-BI over multi-decadal periods.

Furthermore, the SOM-BI can provide unique dynamical information, e.g. concerning spatial information about the location of the blocking anomaly. New sections have been added to the manuscript to describe how the SOM-BI can provide this information (section 2.7), and an application of the SOM-BI for ERA5 has been included in section 3.7.

A discussion of the utility and further application of the SOM-BI is included in the discussion section (L 586-593):

“The use of SOMs as a blocking index provides opportunities for regional study that are not directly available in the other BIs. Through an additional post-processing step involving K-means clustering on blocked node groups (sections 2.7 and 3.7), we have shown that the SOM-BI can identify specific types of blocking events and provide detailed information about the changing nature of blocked events over a European subdomain (Fig. 12). The case of $k = 4$ has been shown in Fig. 12, but larger values of k can also be chosen to identify more distinct types of blocking pattern. Whilst the SOM-BI does not directly produce a gridded climatology of blocking patterns, we have shown that the SOM-BI can be integrated with the other BIs to develop a climatology that only considers only those days detected as blocking by the SOM-BI. This results in a SOM-BI climatology with a higher precision than the BI climatology.

We intend to apply this method to future trends across CMIP5 and CMIP6 models to better understand the patterns of blocking in models, diagnose model skill at reproducing the historic patterns of European circulation regimes and compare projections of future changes in blocking patterns. The identification of distinct blocking patterns from node groups enables a detailed study of blocking characteristics over European subdomains as

shown in Fig. 12. Further quantities such as the Rossby wave breaking properties or the nature of blocking onset and decay can also be studied. This could be done by studying particular dynamical quantities on the blocked days identified by the SOM-BI, and extended by contrasting the dynamical quantities across different categories of blocking pattern identified by the SOM-BI node groups."

The sinuosity index seems a bit disconnected from the other indices since it's not a direct measure of blockings. I would advise to remove this part of the analysis from the article, especially since it's a hemispheric measure and your study applies to regional blocking indices.

The discussion of sinuosity has been removed and replaced with a discussion of K-means clustering, also in response to Reviewer #1.

I am not a specialist of SOMs but I am more used to k-means algorithm and weather regimes. You never cite this part of the literature but I think it would be important to do so since that's something familiar for a large part of the atmospheric dynamics community.

Michelangeli, Paul-Antoine, Robert Vautard, and Bernard Legras. "Weather regimes: Recurrence and quasi stationarity." *Journal of Atmospheric Sciences* 52.8 (1995): 1237-1256.

Vautard, Robert. "Multiple weather regimes over the North Atlantic: Analysis of precursors and successors." *Monthly weather review* 118.10 (1990): 2056-2081.

It would be good to discuss the differences between your approach and what is done in the context of weather regimes.

We have removed the sinuosity discussion and sinuosity in the figures and replaced it with K-means clustering for the case of k=4 has replaced sinuosity in Figures 5, 6, A1, and A2. In L 209-212 some of the literature that uses K-means clustering to study mid-latitude variability has been cited and briefly discussed, including the references that you suggest:

"K-means clustering analysis (Diday and Simon, 1980) has also been extensively used to study the Euro-Atlantic midlatitude variability and identify weather regimes (Vautard, 1990; Michelangeli et al., 1995; Cassou, 2008; Ullmann et al., 2014; Strommen et al., 2019; Fabiano et al., 2021)."

The comparison in the case studies between SOMs and K-means analysis has also replaced the sinuosity discussion in L 350-357:

"Figures 6d and 7d show a K-means clustering analysis using Z500 anomaly fields for the case of 4 centroids. As described in section 2.4, the case of K-means with 4 centroids produces a similar set of weather regimes to SOMs with 4 nodes. Consequently, the K-means analysis exhibits a similar behaviour to the SOMs discussed above but distinguishing between fewer weather regimes. One weather regime indicating Scandinavian blocking consistently represents the 2003 European heat wave across Fig. 6d, but the Westward shift of the high pressure centre from Scandinavia on 31 July to the UK on 8 August 2003 is not described by 4 centroids. For the 2019 heat wave in Fig. 7d, all four weather regimes are represented, and the blocked period is primarily associated

with a mixed weather regime. This shows that the 2019 case is also not described well by K-means clustering.”

The key difference between Self-Organizing Maps (SOMs) and K-Means clustering is that SOMs use a neighbourhood function, which for each stage of the update enables neighbouring nodes to shift towards the best matching node (Figure 3). As a result, in each step not only the best matching node (equivalent to the closest centroid in k-means by e.g. Euclidean distance) is updated, but also – to a degree – neighbouring nodes on the map. Consequently, with increasing node number, SOMs represent a smooth topology across the nodes from left to right and top to bottom on the final map, a realistic continuum of weather patterns. However, for small centroid/node numbers, the difference between K-means and SOMs is small, because the SOMs will still differentiate the range of possible weather patterns, but within a few nodes, making a smooth transition across the map of nodes impossible. Consequently, the results for SOMs and K-means approach each other for small numbers of nodes (n)/centroids (k), as shown in the revised manuscript in Figure A4 for the case of k and n equalling 4. This figure is also shown below. Therefore if K-means for $k=4$ was used as a blocking index in the same way as done for SOMs in our manuscript, it would give similar results to those shown in Figure 7 for $n=4$.

However, this difference between K-means and SOMs is large for high node numbers. We have shown from Figure 7 that the optimum node number is 20 nodes for the SOM. With $k=20$, the K-means algorithm will lead to a set of distinct weather regimes, whereas the SOM will create a continuum of weather patterns. The SOM-BI associates node groups with blocking patterns which is consistent with the behaviour of the SOM as it represents a smooth topology.

A comparison of the SOM nodes for and the K-means nodes has been attached to show the similarity for low node numbers. In L 245-252 these differences have been highlighted:

“This property of SOMs is also the distinguishing feature between SOMs and K-means clustering. In the case of K-means clustering, each node is updated independently and no neighbourhood function is applied. K-means tries to maximize differences between the centroids such that it does not learn a topology. This difference between K-means and SOMs is minor for low node numbers, since the sharp differences in spatial patterns are imposed on the SOMs and the neighbourhood function has a limited effect. For larger node numbers, the SOM topology becomes smoother and the K-means centroids remain distinct rather than representing a continuum of states, whereas a continuum is a more realistic reflection of the actual atmosphere (Skific and Francis, 2012). A comparison between SOMs and K-means analysis for 4 and 20 node/cluster numbers is shown in Fig. A4.”

Specific comments:

The abstract is a bit misleading. The SOM-BI doesn't work well with the 2019 case and this should be acknowledged.

The abstract has been modified in L12-15 to acknowledge the identified limitation of the SOM-BI:

“We present case studies of the 2003 and 2019 European heat waves and highlight that well-defined groups of SOM nodes can be an effective tool to diagnose such weather

events, although the domain-based approach can still lead to errors in the identification of certain events in a fashion similar to the other BIs.”

L. 58 do you mean objective or subjective?

The term “objective” has often been used to contrast automated procedures for blocking identification and subjective human analysis e.g. Tibaldi and Molteni (1990). It has been removed here to avoid misunderstandings.

L.254 I think there's a mistake with R=0.03. Shouldn't it increase compared with 0.19?

We list the skill of each node group individually here rather than the skill of the set of node groups being summed together (which is probably how the reviewer interpreted it). Consequently, P can be lower for a node group while R does not necessarily need to increase accordingly – it really depends on the characteristics of the specific node group. The wording has been clarified accordingly in L289-292:

“For the 3x3 SOM learned from ERA5 data, the node group with the highest precision is [1], with $P = 0.91$ and $R = 0.15$, followed by [2] with $P = 0.89$ and $R = 0.19$ and [1, 2, 6] with $P = 0.87$ and $R = 0.03$. If only one node group is included in the set (e.g. [1] or [1, 2, 6]), there is a high P and low R, but as more node groups are added to the set of node groups ([1], [2]; then [1], [2], [1, 2, 6]), P decreases but R increases.”

L.263 remove « be »

Thank you. This word has been removed.

L.309 rephrase « are not more generally »

Thank you. This has been rephrased to “are not generally”.

Figure 7 and 8 look very flat to me. Since those are the curves you use to determine your optimal number of nodes I would discuss this in the paper and potentially introduce some sort of confidence interval around your optimal k.

Whilst the dependence of the F1 score appears quite flat, we show in Figures 7 and 8 that we also consider the balance of precision and recall and not just the F1 score. The procedure we use for identifying the optimal number of nodes is discussed in L 439-444:

“Common features are observed for each variable for a very small or large number of SOM nodes. For small k the SOM-BI identifies more days as blocked, such that $R \gg P$. This indicates that the SOM is under-fitting the data for European circulation patterns across the domain and so the algorithm lacks a precise delineation of blocking events. In other words, it could be beneficial to increase k to be able to represent a larger number of dynamical states and thus to detect and describe blocking events more precisely. For large k, $R \ll P$, showing that the SOM-BI is trending towards overfitting the training data. We deduce that the optimal k occurs when the difference between P and R is small and the F1 score is close to its maximum value.”

We consider the minimum difference between precision and recall and the peak of F1 score, which together lead to a close approximate value for the number of nodes. We think

that this procedure quite clearly outlines a way for choosing a consistent number of nodes in any given case, e.g. of SOM-BI was applied to a different domain. We would thus rather refrain from defining an ambivalent confidence interval that could confuse this message for the general reader.

L.419 and 421 I think you're talking about Figure 10 here, not Figure 9.

Indeed – thank you. This Figure label has been corrected accordingly.

Response to Reviewer #1 minor revisions

The authors greatly improved the manuscript addressing the most of my comments with complete and accurate replies. I have just a few minor concerns that I list here below:

L23: You said that your method is an unsupervised learning: although this is true for SOM, however you exploit a ground truth to define which node found by SOM is blocked or not. Would be more correct to call it as a supervised learning?

It is true that our method uses supervised learning through using the ground truth to identify the SOM node groups associated with blocking. However, since the nodes themselves are identified through the SOM, we think it is more helpful for the reader to describe our approach as unsupervised, to highlight the contrast between our method and the conventional blocking indices.

L63: Even if I found it fun to read, I wonder if the pronunciation is really necessary...

This pronunciation is not necessary, and it is primarily meant to help the algorithm stick in the reader's mind.

L78: what do you mean with skill here? Please rephrase.

Skill here refers to blocking identification skill, F1 score in particular, which is defined in section 2.6. We have therefore rephrased it to “blocking identification” and changed L78 sentence from:

“We identify a moderate improvement in blocking identification over the BIs for the reanalysis period and a significant improvement in skill for the UKESM data.”

To:

*“We identify a moderate improvement in blocking identification over the BIs for the reanalysis period and a significant improvement in **blocking identification** for the UKESM data.”*

L91-109: the paper is already quite long, the outline of the method section could be removed.

The outline of the method section has been added to help the reader navigate the lengthy manuscript.

L121: relative vorticity?

Correct. "Relative" has been added to L121.

L164: I would remove Figure 2. It is true that you analyzed MSLP instead of PV, but the approach to define the GTD is the same.

It is true that Figure 2 shows the same process of labelling with a minor difference of variable, but it also shows the labelling procedure in a climate modelling context so we still think that it should be included in the manuscript. We have therefore moved the figure into supplementary information as Figure A1.

L215: Why fifth? I count three blocking indices, shouldn't this be the fourth?

Correct. L215 has been amended.

Figure 3: Please write the BMU acronym in the caption. The 10^{-3} multiplication is unclear: do you mean that the initial SOM anomaly has a +270000 meter anomaly?

The BMU acronym is defined in step 2 in the figure, but we have added a definition in the caption as well. The initialised pattern in step 1 does indeed have very large anomalies. The caption of Figure 3 includes a clarifying statement. The caption has changed from:

"The self-organizing map algorithm. Shown using a 3x3 node SOM with ERA5 Z500 JJA 1979-2019. The initialised SOM pattern (step 1) has a much larger amplitude so has been multiplied by 10^{-3} for visualisation purposes."

To

"The self-organizing map algorithm. Shown using a 3x3 node SOM with ERA5 Z500 JJA 1979-2019. The PCA-initialised SOM pattern (step 1) has a much larger amplitude so has been multiplied by 10^{-3} for visualisation purposes. The BMU refers to the best-matching unit, the SOM node which most closely matches the sample day."

L265: here and in other instances you refer to "5-day period". However blocking events in the GTD can be longer, is this correct? Therefore why not calling them blocking events or something like that? More in general, although some clarification in the new version has been presented I still find confusing the methodology for the blocking definition in the GTD. For example, at L268 you say: "Since each five-day period has been classified either as blocked or not blocked, [...]" This sentence makes no sense to me, since EACH DAY should be defined as blocked or non-blocked (even if this definition is based on a 5-day threshold). It seems to me that here you are trying to identify which node describes each blocking event, but blocking events last more than five days.

It is true that blocking events frequently last more than five days. This is reflected in our GTD, which labels each day as blocked or non-blocked. However, since a blocking event cannot last for less than five days (a common persistence criteria) we need to study groups of five days to identify for each individual day if a day is blocked. One cannot identify a block by studying a map for a single day in isolation.

By studying every consecutive five-day period across the dataset, we can see if a block persists across that period or not. This will identify every blocking event across the dataset.

Blocking events longer than five days are identified in this way from the identification of overlapping five -day periods as blocked.

For example, if two consecutive five-day periods are labelled as blocked (across days 1-5 and days 2-6), then there is a blocking event that lasts six days (across days 1-6 inclusive). If fifteen consecutive five-day periods are labelled as blocked, then a blocking event of 20 days in duration has been identified.

Therefore, by studying every five-day period we identify all blocking patterns across the ground truth dataset.

We have amended the explanation by adding a sentence in L162:

"Blocking events longer than five days are also identified through this approach, since consecutive five-day periods will be labelled as blocked."

L381: I still feel this approach is overly inclusive for blocking events. The climatology you get from Table1 is quite unrealistic, with AGP having 60% of blocked days, twice than DG83 – which is usually detecting much more areas as blocked (see Fig 6-7) and have larger climatological frequencies than AGP (Fig 2 in Woollings et al 2018). Usually spatial and temporal criteria are applied to blocking indices to avoid detection of spurious events, and this does not seem to have been done here.

The spatial and temporal criteria have been applied across all indices (see below), and the numbers in Table 1 reflect the results that we obtain from our analysis.

Note that there is an important contrast between figure 2 in Woollings et al. 2018 and in the case studies in figures 6 and 7: Table 1 labels a blocked day if there is a block that exists within the European sector and satisfies all of the relevant blocking criteria. It is therefore much more dependent on the *time series* of blocking across the sector than on the scale and magnitude of the identified blocks. We have found that AGP identifies smaller events across more days in the dataset than DG83, which is why the number of events identified by AGP is so large.

We have attached an ipython notebook which shows our analysis to derive the AGP criteria and our application of the area, persistence and overlap criteria, and also how the 58.8% figure is derived.

L382: please reword: "if a blocking event has been identified" -> "if a single grid point has been identified as blocked".

This is a misunderstanding. We do not identify blocking events across Europe on the basis of a single grid point, but if a blocking event has been identified within the European sector. This requires the block to meet all of the various thresholds for persistence, area and overlap. L382 says "if a blocking event has been identified within the European sector. A blocking event is not identified if the thresholds for amplitude, persistence, area and overlap discussed in section 2.3 are not met within the European domain." – these statements accurately reflects our analysis.

L383: which threshold in amplitude, persistence area and overlap are you referring to? In Section 2.3 only the indices are defined, no discussion of the tracking is introduced.

In L 185-190 in section 2.3 the amplitude, persistence, area and overlap thresholds are referred to:

"All of these methods have been further developed by Pinheiro et al. (2019) who applied four thresholds for each blocking index: the magnitude of the anomaly, the persistence of the blocking event, a minimum area over which the anomaly takes place and an overlap criterion which measures if there is continuity across the blocked region between different days. We adopt their thresholds and as such study the three indices compared in Pinheiro et al. (2019) including their modifications."

– AGP - the geopotential height gradient method, which is the Tibaldi and Molteni (1990) index as adapted by Scherrer et al. (2006) to construct a two-dimensional field of geopotential height gradients

– DG83 - the Dole and Gordon (1983) method of investigating positive geopotential height anomalies

– S04 - the Schwierz et al. (2004) method of identifying persistent anomalies in the potential vorticity field (VPV) averaged over 150-500 hPa (VPV).

However, our analysis differs from the methodology..."

We have expanded this section to refer to the specific criteria, and to limit the length of this manuscript we refer the reader to Pinheiro et al. (2019) for an extensive discussion of these criteria:

*"All of these methods have been further developed by Pinheiro et al. (2019) who applied four thresholds for each blocking index: the magnitude of the anomaly, the persistence of the blocking event (**minimum five days**), a minimum area over which the anomaly takes place (**10^6 km^2**) and an overlap criterion which measures if there is continuity across the blocked region between different days (**an overlap of the blocked contours**). We adopt their thresholds and as such study the three indices compared in Pinheiro et al. (2019) including their modifications."*

– AGP - the geopotential height gradient method, which is the Tibaldi and Molteni (1990) index as adapted by Scherrer et al. (2006) to construct a two-dimensional field of geopotential height gradients

– DG83 - the Dole and Gordon (1983) method of investigating positive geopotential height anomalies

– S04 - the Schwierz et al. (2004) method of identifying persistent anomalies in the potential vorticity field (VPV) averaged over 150-500 hPa (VPV).

We refer the reader to section 2.2 in Pinheiro et al. (2019) for a detailed discussion of these methods and their associated thresholds. However, our analysis differs from the methodology..."