Authors' replies to Referee #1

We thank Referee #1 for their constructive comments. Below, we provide our replies; the line numbers and sections refer to the old version of the manuscript.

Major Comments

1. WTD

1a. Discussion on advantages and limitations of WTD. *I find no discussion in the paper about advantages and limitations of WTD (except description of how it is done and that it agrees with blocking indices). One of the limitations of WTD is that WTD is less flexible and it does not come with a native measure of blocking center and size.*

We agree with the Referee that such a discussion is currently not well developed. Following also the suggestions of Referee #2, we discussed pros and cons of the WTD but also of the blocking indexes in the Introduction. Please, find below the new text (in blue) added in paragraph 28-38.

"The identification of blocking events in numerical simulations is complicated by the fact that blocking is determined by various dynamical mechanisms and presents different patterns. Several blocking indexes have been proposed in the literature, based on meteorological fields, usually the geopotential height at 500 hPa (e.g. Tibaldi and Molteni, 1990), or anomalies of meteorological fields (e.g. Dole and Gordon, 1983). Blocking indexes focus on different characteristics of blocking, so the choice of the index depends on the purpose of the study. Additionally, index definitions depend on various (user-dependent) parameters, like latitude band limits, latitude references, and anomaly thresholds (a review of the blocking indexes can be found in Barriopedro et al. (2010), while a recent discussion about their differences is in Pinheiro et al. (2019)). Given the variety of blocking indexes, the comparison across studies is not straightforward.

Atmospheric blocking can also be identified via the so-called weather type decomposition (WTD) methodology, which classifies the atmospheric circulation into discrete weather regimes (Michelangeli et al., 1995). The WTD methodology, referred to as *the WTD* hereafter for brevity, relies on a partitioning algorithm that groups data of a meteorological variable (usually geopotential height or sea level pressure) into clusters so that the variance between clusters is maximized and the variance within a given cluster is minimized. In this way, the clusters (weather regimes or weather types) are the result of a mathematical algorithm. The results of the WTD depend on certain user choices, such as the sector size, the clustering algorithm and the initialization of this algorithm. Despite the fact that the clusters may not be well separated, WTD has proved to be very useful in the literature. In fact, WTD allows to explain most of the atmospheric variability and has largely been used to define weather regimes especially in the Northern Hemisphere (e.g. Michelangeli et al., 1995; Cassou et al., 2004; Barriopedro et al., 2006; Ullmann et al., 2014; Fabiano et al., 2020). In the European-Atlantic sector, for example, four winter weather types have been recognized: positive North Atlantic Oscillation (NAO), negative NAO, Atlantic ridge, and European blocking. The WTD has also been used to analyze weather types in relation to other quantities like temperature (e.g. Cassou et al., 2005), precipitations (e.g. Ullmann et al., 2014), winds (e.g. Jiménez et al., 2009), and

pollutants (e.g. Russo et al., 2014). In this study, the WTD is used to identify blocking events in the European-Atlantic sector."

1b. Fitness of WTD. *I wonder how well WTD can summarize the Z500 variability. In the process of k-mean, is variance between cluster much larger than variance within the same cluster? Or do the 4 clusters explain a very high percentage of variance? Or is the clustering very clear cut?*

There is a consistent number of studies showing that the WTD applied on Z500 can be used to divide the atmosphere at the synoptic scale into weather regimes, see the references cited in the Introduction and the new text written in point 1a above. Most of the studies considering the European-Atlantic sector in winter show that the clustering is optimized with k=4, so the WTD used to define the four weather regimes in this sector can be as a "standard procedure". The evaluation of the WTD methodology is beyond the purpose of this work, rather we evaluated the WTD results by comparing the four weather types with previous literature (section 4.1); since they agreed well with previous studies, we could develop our analysis on the basis of these WTD results. In the new version of the manuscript, we added a new Figure (please, see Fig. 1 in the replies to Referee #2) to show the agreement between the blocking episodes identified via the WTD and the DG-index.

1c. Interpretation of WTD. *If the WTD clustering is not that clear cut, it is hard for me to interpret the results. When "cluster centroid" is found different in some models, what does it mean? Is it because reanalysis-blocking-like pattern occurs less frequent? Or is it because some boundary cases (under reanalysis clustering) occur more frequent and that the cluster boundary needs to be put elsewhere?*
Also the "blocking frequency". Do changes come from the 5-day requirement, or the overall frequency of weather type? If it has to do with the overall frequency of weather type, the question again is whether it comes from changes in frequency of centroid-like patterns or boundary-case-like patterns? Same question for the "blocking center". Does it come from the weather type shifting in location? If so, is it because of centroid-like patterns or boundary-case-like patterns?

Differences of weather types among GCMs occur because global simulations depend on factors like internal climate variability, resolution, orography, and parameterizations that are GCM specific. Thus, there is the possibility that a GCM captures some recurrent patterns less well than other GCMs, affecting the WTD results (i.e. the weather types resulting from the clustering). The evaluation of the models in subsection 4.1 has exactly the aim of selecting those GCMs that well reproduce the weather regime of blocking, following the approach of other studies (cited in 4.1 subsection).

The frequency of the blocking days resulting from the clustering differs a bit (about 20%) from the frequency of the blocking days belonging to the WTD-blocking events. This is natural given the application of the criteria defined in subsection 3.2, e.g. the minimal duration of 5 days to define a blocking event. We would like to point out that such criteria are the ones used in the literature when blocking days are identified via blocking indexes, thus, we expect that also in that case the frequency of the blocking days is slightly different from the frequency of the blocking days belonging to the blocking events (although, we are not aware of any

quantification of this "frequency change" in the literature).

1d. Insignificant results. *If the WTD clustering is not that clear cut, I wonder if this will cause extra variability that stops you from drawing significant results. Most results using DeltaZ500_SSP are not statistically significant.*

We do not know if this "extra variability" is the cause of getting not significant results. We cannot exclude this, but we would like to point out that one could obtain not statistically significant results also with a clear cut of the clusters. Finding not statistically significant changes is a result itself, and this was the case also for other works cited in the manuscript (in fact, the review of Woollings2018 states that no clear long-term changes have been found in blocking frequency).

2. DeltaZ500_HIST may be irrelevant.

Many results are based on DeltaZ500_HIST, in which the overall higher geopotential height in warmer climate is not removed. I cannot see how this overall higher geopotential height would link to weather impact or air pollution, which is likely why authors are interested in blocking. Measures of blocking based on DeltaZ500_HIST go too far, and become irrelevant to weather impacts.

We agree with the Referee that if we analysed climate change impact on air pollution during blocking we should consider the results obtained with DeltaZ500_SSP. Following also the suggestions of Referee #2, we decided to present the DeltaZ500_SSP results in the revised version of the manuscript and to move most of the DeltaZ500_HIST results to the Supplement (or to remove them).

Minor comments

3. Line 103: How do authors determine which (out of 4) weather regime is the European blocking weather regime?

The four weather regimes obtained via the WTD in this study are close and comparable to the (usual) four weather regimes of the European-Atlantic sector described in previous literature. For instance, Figure 1 (below) shows the four weather types obtained with the ERA5 reanalysis in this study. We can observe that the association between the weather types (WTs) meant as the centroids resulting from k-means and the four atmospheric weather regimes is clear: WT1 is blocking, WT2 is Atlantic ridge, WT3 is NAO+, and WT4 is NAO-. Since the WTD was applied to each model, such association was defined each time.

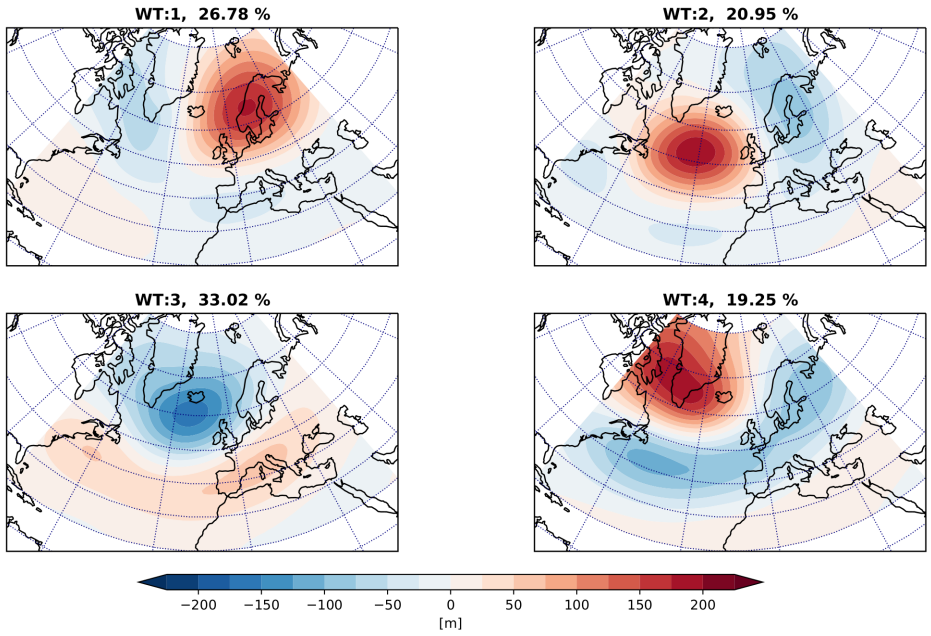

Figure 1: The four weather regimes obtained with the WTD (and k = 4) over the European-Atlantic sector using ERA5 reanalysis (1980-2009).

4. I prefer to say "size" or "area" of blocking, in place of "extension". The latter is not clear to me and let me think of temporal extension (duration), or some kind of extension of concept.

We replaced "extension" with "size". The term "area" is used to quantify the blocking size, as written at line 129.

5a. Line 131: I suggest change "center method" to "composite method". Because "composite" is really the step that differs from the "DG method".

Ok, we replaced "center method" with "composite method".

5b. But actually, the authors made a few modifications to the DG method that makes it very similar to the center method. For example, authors require DG blocking day to be a subset of WTD blocking day. Perhaps authors need to say what are the major remaining differences (if any) in the two method.

We did not modify the procedure followed by Nabizadeh et al. 2019, a part from the fact that it is applied on WTD-blocking events (as the Referee pointed out), and we would not say that the two methods are similar. Although the algorithm to compute the blocking area is the same, it is applied on composites in the composite method and on daily anomalies in the DG-method. Moreover, the ΔZ500 values of the contour lines (i.e. the thresholds) are defined in different ways: they depend on the center intensities in the composite method, while they depend on $\widetilde{\sigma}_{max}$ in the DG method. We made a few modifications in the Supplement (section "Computation of blocking area") to improve the description of the two methods and we highlighted the differences between the two methods at line 147:

"Therefore, these methods compute the area of blocking events that are identified via two

different approaches: the WTD and the DG index. Although the algorithm to compute the blocking area is the same, it is applied on blocking composites in the composite method and on daily ΔZ500 in the DG method. Another difference between these two methods is the definition of the ΔZ500 values of the contour lines (i.e. the thresholds)."

6. Line 7: I think your methodology to quantify size of blocking does not "rely on the WTD". I don't think it is WTD-native or WTD-specific. I don't think this is sufficiently different from other studies (like Nabizadeh et al. 2019) that I would claim "new".

We removed the adjective "new". We preferred to leave "rely on the WTD" as the center method is applied on WTD-blocking events, moreover, since we changed "center method" to "composite method" (as asked by the Referee), we think that this expression ("rely on") could remain in the Abstract.

7. Line 9: "Geopotential height increase" might be more accurate than "pressure increase".

Done.

8. Explanations of methods are disordered. For example, the 3 paragraphs in section 3.1 go by talking first about WTD, then an overview of all steps, and lastly the calculation of Z500 anomaly (which is done before WTD).

We changed the first two paragraphs in subsection 3.1 in order to explain the methodology step by step. We moved the third paragraph in a new subsection with title "Z500 anomalies" to explain the meaning of ΔZ500.

9. Line 105: I would suggest to add a bracket "(including the mean)" after "annual cycle". Because "annual cycle" can sometimes only refer to the seasonal variation from the mean.

We made this addition at line 97, where we write "annual cycle" for the first time.

10. Line 119: I am not sure the description on treatment of "hole" is complete that others can reproduce. Let 0 be non-blocking and 1 be blocking. What would the code say about 001010100, and 001110101011100?

The code processes the "labels" resulting from k-means (i.e. 0, 1, 2, 3 for k=4) in the following order: 1) two blocking events longer than two days separated by a hole form one blocking event; 2) one blocking day and one blocking event longer than three days separated by a hole form one blocking event (the code checks first the case with order 11101 and then 10111). The code stops searching for the holes after checking the conditions 1 and 2.
The first example reported by the Referee (*001010100*) does not satisfy the previous conditions, therefore, all zeros remain holes. The second example (*001110101011100*) satisfies the second condition: *001110101011100* → *00111* ***1*** *101011100* → *001111101* ***1*** *11100*. Therefore, the final result is: 2 blocking events of 5-day duration. We would like to reassure that these examples are extreme cases, as usually the k-means result (4530 labels, i.e. the number of days of 30 winters) does not present such an "unstable" sequence.
We modified the text to be more precise in the explanation at line 121:

"Therefore, the k-means result is processed in such a way that 1) two blocking events equal to/longer than two days separated by a hole form one blocking event and 2) one blocking event equal to/longer than three days and one blocking day separated by a hole (and then vice versa) form one blocking event."

11. Line 125: Is "hole" included in "blocking days"?

We call blocking days only those days which belong to a blocking event (which is defined in subsection 3.2). After the identification of blocking events, we do not speak about "holes" anymore but only about blocking days, although a blocking day could be an "old" hole.

12. Line 136: I prefer a simpler phrase "non-zero" in place of "non vanishing".

Done.

13. Line 136: I would suggest to mention the 75m/100m threshold here in main text, rather than having to find it in supplement.

We added the threshold values in subsection 3.3.

14. Line 154: With Fig. 1, what is being evaluated is not ability to reproduce the "blocking weather regime" but "composites" (as defined in line 124-127).

Done.

15. Fig. 1: I assume this figure is based on DeltaZ500_HIST, so the overall higher geopotential height is included? From Fig. 3, I guess the overall Z500 increase is more than 25m in SSP2. Why don't I see an increase of RMSD because of this?

There is an inaccuracy in the text, we thank the Referee to point it out. The Taylor diagram shows the central root-mean-square difference (CRMSD $= \sqrt{\frac{1}{N}\sum_i \left[(p_i - \bar{p}) - (r_i - \bar{r})\right]^2}$, where p = prediction (GCMs) and r = reference (ERA5)), therefore, the effect of the overall higher Z500 is not included in the results for SSP2 and SSP5. If we computed the usual RMSD for $\Delta Z500_{HIST}$ we would obtain, for example for MPI, 48 m in SSP2 and 82 m in SSP5, while the CRMSD values (in Fig.1) are: 18 m in SSP2 and 29 m in SSP5. The latter values are almost equal to the RMSD obtained for $\Delta Z00_{SSP}$. In the revised version of the manuscript, we showed only the results obtained with $\Delta Z500_{SSP}$ in the Taylor diagram and computed the usual RMSD.

16. Table 1: The resolution of GFDL is said to be 1 degree on https://wcrp-cmip.github.io/CMIP6_CVs/docs/ Could the authors please check? I assume the argument made in line 179 is based on the resolution when the model is run, not the resolution of the output.

We checked the source files stored in the Mesocentre ESPRI (in /bdd/CMIP6/CMIP/NOAA-GFDL/GFDL-CM4/historical/r1i1p1f1/day/zg/gr2/latest/) and we can confirm that the resolution of GFDL-CM4 is 2.5°x2.0° (as written in Table 1), with 12960 grid points for the globe (144 in longitude x 90 in latitude).

Yes, the sentence at line 179 refers to the model runs.

17. Line 211/227: I am not sure the similarity between DeltaZ500_HIST and DeltaZ500_SSP is entirely interesting. The overall increase in geopotential height is a shift of all clusters in a hyper space. So by construction, it has no effect on the clustering result. The only difference is the seasonal variation around the mean. The similarity in results can only suggest the seasonal cycle does not alter enough from HIST to SSP to alter the clustering result.

As mentioned before (at page 3), we showed the DeltaZ500_SSP results in the revised

version of manuscript and we moved most of the DeltaZ500_HIST results to the Supplement (or we removed them).

18. Fig. 5: The peak of ERA5 at 27-28 days look suspicious.

We checked the computations and the plot is correct. We obtain 1 blocking event of 27 days and 2 blocking events of 28 days (so we have three events in the 27-28 bin in Fig.5). This looks less suspicious if we look at the plots in Fig.S5: events that are longer than 20 days occur rarely in 30 years but it can happen to see a peak, like also for GFDL in HIST and for INM in SSP2.

19. Line 247: Perhaps you can clarify "variability". Do you mean inter-model variability, or inter-event variability?

We meant "inter-event" variability (however, this sentence is not present in the new version of the manuscript).

20. Line 267: I don't think the 0.1% is a significant digit if the area only has two significant digits.

Ok, we removed the percentage.

21. Line 270: I think Nabizadeh et al. 2019 is based on DeltaZ500_SSP. And the increase you are talking is drastically larger than 17% in Nabizadeh et al. 2019. I don't know if I would call this agreement.

This sentence is not present in the new version of the manuscript.

22. Line 304-305: The sentence looks contradicting to me. ("may not match", "agrees well").

We modified the sentence to "Despite the number of DG-blocking days may not match with the duration of the WTD-blocking events, we find that it generally agrees with the duration of the WTD-blocking events." This concept should be clear in the revised version thanks to the addition of the Figure included in the reply to Referee #2.

23. Supplement Step A: I assume this step applies both to the center method and the DG method. But the step uses "blocking center". What is the "blocking center" for the DG method? Also, for the DG method, is there at most one such center/blob on each day, such that step 8 in the DG method only does temporal mean but not event mean?

We thank the Referee to point this imprecision out, in fact, Step A applies to both methods but "blocking center" was defined only for the center method. In the revised version of the manuscript, we called "DG-grid boxes" those grid boxes where DG > 1.5 for at least five consecutive days and we corrected the explanation in Step A.

In the DG method there could be more than one blob per day, therefore, only the blob including at least one DG-grid box defines the blocking area of that DG-blocking day (as explained in Step A). The temporal mean of the areas of the DG-blocking days belonging to the same WTD-blocking event determines the area of the DG-blocking event.