Dear Editor,

thank you for your comments and suggestions. In this document, first we reply to your two comments, then we report the modifications that we implemented in the new version of the manuscript.

L53-55: it would be helpful to add a few more words and mention that Hassanzadeh et al's paper showed a decrease in blocking activity with Arctic warming or just that it highlighted the need for more study on this link, or something along these lines.

We modified the sentence; please, see our reply to Referee #2.

Please carefully address the reviewer's comment 1 and their other comments/questions. You might also want to take a look at this paper by Falkena et al. https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.3818. Based on these comments and these studies, I suggest that you consider taking a look at the potential effects of increasing k (e.g., to 6).

We thank the Editor for this interesting reference, which we cited in the manuscript. Falkena et al. (2020) found that k=4 is the best choice when k-means is applied on the reduced data, while they showed that k=6 is the optimal choice when k-means is applied on the full field data. Since it is well documented that k=4 is a good choice when k-means is applied after PCA, we did not repeat the analysis for k=6. Please, see our discussion about k in the reply to comment-1 of Referee #1.

The major changes in the last submitted manuscript are in the Conclusion section (now named "Conclusions and discussion"), where we added some discussions and comments about the points raised by the referees. Small changes are present in the Introduction and in the presentation of the results (sections 4.3 and 4.4), as we removed two plots regarding the $\Delta Z500_{SSP2-HIST}$ and $\Delta Z500_{SSP5-HIST}$ results and reorganized the text and the figures.
These changes and all the smaller corrections are visible in the marked-up version of the manuscript.

Thank you very much!

Yours faithfully,

Sara Bacer

Authors' replies to Referee #1

We really thank the referee for the interesting comments; they were helpful to improve the manuscript. Below, we provide our replies.

Major comments

*1. **WTD.** After my first review, I came across Dorrington and Strommen 2020 (hereafter DS2020) and Dorrington et al. 2021, which answered some of my questions about WTD.*

Fitness of WTD: DS2020 found that the raw Z500 phase space (before removing the jet speed variability) is quite Gaussian (DS2020's Fig 1a). In other words, the clustering is not very clear cut. DS2020 found the standard/classical k=4 clustering to be problematic. "When setting K=4 in our residual space, we found k-means clustering consistently returns Clusters 3, 4, and 5 of Figure 4 but in different 30-year windows switches between including Clusters 1 and 2" (quoting from last paragraph of section 3.4 in DS2020). Such decadal variability in cluster centroid was considered not a real signal, but artifact of the choice of k=4.

*Interpretation of WTD: Back to your work, your response to my previous comment 1c did not address my questions. (Maybe I was not clear enough.) I still cannot understand what is different in the *raw model output*, that causes difference in the WTD analysis. Part of my questions relate to your choice to allow cluster centroids to differ and that the clustering is not clear cut. DS2020 made the clustering clear cut by removing the jet speed variability, and they worked towards requiring "dynamically relevant regimes to be approximately stationary features of the midlatitude circulation over centennial time scales, at least in terms of spatial patterns if not in residence times or transition probabilities" (quoting from section 3.2 in DS2020).*

When you find "cluster centroid" to be different in some models, could it be the models behave like different 30-year windows in the example of DS2020? And that difference gets overly exaggerated because of the bad choice of k=4?

On the interpretation of "blocking frequency" and "blocking center", I would like to see the output of the WTD clustering (the weather type and its frequency) as an intermediate step before the "blocking frequency" and "blocking center". I think that helps readers interpret the results. Meanwhile, fundamentally, I am looking for more clear-cut clustering, or to use the same cluster centroid. In this way, I can better trust that WTD clustering can faithfully summarize the Z500 variability, and will not overly exaggerate noises.

We thank the referee for these interesting recent references.

We would like to point out the following considerations.

- In DS2020, the analysis is repeated for different 30-year windows along the 20th century, e.g. Fig.2b. Since their aim is to perform a statistical study of criteria such as the BIC, the analysis does not show the results for each 30-year period explicitly. Therefore, we cannot know if the results for the period 1902-1931, for instance, are close or not to the results for the period 1912-1934 (i.e. shifted by 10 years).
- In our study, instead, we consider only two 30-year periods in order to study the difference between the end of the 21th century and the past.

- There is no rule about how many years to consider for the WTD; for instance, Cassou 2008 considered 33 years and Ullmann et al. 2014 considered 26 years.
- If we considered 40 years or other 30 years shifted by 10 years (e.g. 1970-1999 for the past and 2060-2089 for the future) and if we obtained results slightly different from the ones in the manuscript (but still not significant changes between past and future), this would mean that blocking changes are small and within the climate intra- and inter-decadal variability, like suggested in Huguenin et al. (2020). We added some comments at L320 in the conclusions (please, see the end of this reply).
- The differences that we find in terms of frequency and persistence of weather types across the models are not new and have already been discussed (e.g. Huguenin at al. 2020).

Therefore, we do not expect that the results of our analysis depend on the selected 30-years.

Considering the published literature, we do not think that "k=4 is a bad choice". k=4 has been widely documented to be the best choice for the classification of winter weather types in the Atlantic sector, and such value of k has been largely used in other studies (please, see the citations in section 3.1 in the manuscript and references therein). It is true that, recently, some studies suggested other values of k, but these findings must be contextualized (e.g. on which data set exactly is k-means applied?) and confirmed by further studies. We write below two examples.

- Falkena et al. 2020 found that k=6 is the optimal choice when clustering is applied on the full field data. However, when the clustering is applied on the reduced data, i.e. the PC data, they found that the optimal number of clusters is 4.
- DS2020 found that k=2, 5, or 6 is a better choice using raw PCs, while k=3, 5, or 6 is better when using residual PCs, i.e. after removing from the geopotential height the influence of the jet speed. The first result is clearly in contrast with a large number of previous studies (all those for which the best choice is k=4), but we could not find any explanation for this disagreement in the paper.
  Moreover, we observed that DS2020 retained only 10 PCs to explain about 83% of the variance. However, the number of PCs can influence the choice of the best value for K. Indeed, Falkena et al. (2020) showed that "*For lower numbers of EOFs, the optimal number is found to be lower, while a higher number of EOFs leads to a higher optimum for k. This is to be expected because the use of a limited number of EOFs means that some variability of the original data is neglected. This loss of variability is larger when fewer EOFs are used, and as a consequence fewer clusters are needed to account for the variability of the EOF data.*" In particular, Falkena et al. found that k=4 is the best choice using 20 PCs, K<4 using 10 PCs.

Overall, we believe that, for our paper, k=4 is not a bad choice and is a recommended one as it is well justified in the literature.

We added at the end of this document the four centroids (for the four weather types: blocking, Atlantic ridge, NAO+, and NAO-) and the frequency of occurrence obtained for all models (for the historical period).

As commented in our replies in the previous report, k-means does not perform a clear-cut clustering. However, our methodology relies on a well established and largely used approach to define weather types: PCA+clustering. Moreover, we find similar results for the size of the blocking area by applying the WTD method and the DG method (which uses the DG index on the raw data), see Figure S6. This suggests that the noise associated with the non clear-cut clustering does not dominate the results.
Finally, we would like to add that, before adopting this procedure, we followed the approach of using reference eigenvectors and reference centroids (i.e. the ones obtained with the reanalysis) to find the weather types of the GCMs (like in Ullmann et al. 2014). However, the blocking pattern obtained with some GCMs did not resemble the reference blocking pattern; instead, by using the PCA + k-means approach for each GCM, we have always obtained blocking regimes close to the reference blocking pattern.

L320 (Conclusions): "Given the decadal variability of weather regimes (Dorrington et al. (2020), longer past and future periods could be considered (e.g. periods of 50 years, like in Fabiano et al. 2020) so as to better smooth the dependency of the results on this decadal variability. Moreover, those days for which the geopotential height anomaly field does not resemble the blocking weather regime pattern could be classified as "neutral days", like in Dorrington et al. (2021), and excluded from the analysis."

L324 (Conclusions): "The optimal number of clusters depends also on the data to be processed; for instance, by applying the clustering on the full field data, Falkena et al. (2020) found that k=6 is an optimal choice."

Minor comments

2. Insignificant results. Related to my previous comment 1d. Now, after taking away results using DeltaZ500_HIST, most results are not statistically significant. If you cannot exclude that non-clear-cut clustering might have a role in making the results insignificant, I think you should acknowledge that in your manuscript.
While I agree that "Finding not statistically significant changes is a result itself", I encourage authors to cite papers that contrast with the results here. Huguenin et al. 2020 also used some kind of circulation type classification, and found lack of change in frequency and persistence, but they cited papers which contrast with their results. You may also read Kautz et al. 2021 and Nabizadeh et al. 2021, where you may find some discussion on blocking under climate change.

We added at L320 (Conclusions) the following sentence: "Finally, a sensitivity analysis of the results to the clear-cut character of the clusters could be conducted, for instance by removing the influence of the jet speed from the geopotential height field, like in Dorrington and Strommen (2020)."

We thank the referee for these interesting references. We changed/added some text in the Introduction and Conclusion (this section is called "Conclusions and discussion" now) to improve our discussion on the impact of climate change on blocking frequency, duration, and size and compare better our results with previous findings (which can agree or not with ours). (Blue parts are the main changes.)

L55-64 (Introduction): "So far, studies have mainly focused on frequency and duration of future blocking events. Some of these studies found that blocking frequency will decrease in the Northern Hemisphere (e.g. Dunn-Sigouin and Son, 2013; Matsueda and Endo, 2017; Fabiano et al., 2020, Davini et al. 2020), while blocking duration may either increase (Sillmann et a. 2009) or decrease (Fabiano et al. 2020). Other studies showed that blocking frequency and duration will not change notably in warming climate (Dunn-Sigouin and Son, 2013; Huguenin et al. 2020). Future changes in blocking size have received less attention (Hassanzadeh et al. 2014; Nabizadeh et al., 2019).

Most of the studies mentioned above determined blocking events via blocking indexes (Sillmann and Croci-Maspoli, 2009; Dunn-Sigouin and Son, 2013; Hassanzadeh et al., 2014; Matsueda and Endo, 2017; Nabizadeh et al., 2019; Davini and D'Andrea, 2020) and considered GCMs participating in the Coupled Model Intercomparison Project phase 5 (CMIP5 Dunn-Sigouin and Son, 2013; Matsueda and Endo, 2017; Huguenin et al., 2020) or idealized GCMs (Hassanzadeh et al., 2014). To our knowledge, only Fabiano et al. (2020) employed CMIP6 models in order to project the blocking weather type and analyse its changes in frequency and duration in the 21st century."

L304-317 (Conclusions and discussion): Please, see directly the new version of the manuscript.

3. *Line 14-15: You may also refer to Kautz et al. 2021 [doi:10.5194/wcd-2021-56].*
We added this citation.

4. *Line 124: Consider change "net impact" to "total impact" or "gross impact".*
We used the expression "gross impact".

5. *Section 3.3: Your response to my previous comment 10 helps a bit, but the revised manuscript is still not clear on the treatment of holes.*
Line 128. "longer than five days"->"at least five days long"
Line 129. Remove "and separated by at least two non-blocking days". Because this is not true for the 2nd example I gave last time (001110101011100).
Line 129. Consider change "is assumed to represent"->"might represent", in order to soften this sentence, because this is not true for the two examples I gave last time.
Line 130. Consider change "Therefore" to "Concretely", because this sentence is what the code does, not only examples.

Below are more subtle details. One way is that you can reply here and refer to the discussion here in the manuscript. If you are doing find-and-replace in place, the searching order of (11011,11101,10111) matters, e.g., 00110110100, 001110101100. Would be good to make explicit the order of searching. Overlapping matches can be bad for codes, e.g., 110111011 is two 11011 overlap together. Does your code find one or two matches of 11011? What will your code say about 0011011101110100?

We changed the sentences at lines 128, 129, and 130, as suggested by the referee.

First of all, we would like to point out an inaccuracy in our reply to the comment-10 of the previous report: the second example provided by the referee in that comment (i.e. *001110101011100*) gives *0011**1**1**1**1**1**11100* (i.e. one blocking event that is 11 days long).

This occurs because our algorithm verifies the conditions (1) and (2) (see L131) considering one by one element of the list of labels, moving from the first element of the list to the last one. Therefore, the sequence of the previous example undergoes the following changes: *001110101011100* $\rightarrow$ *00111* ***1*** *101011100* $\rightarrow$ *00111* ***1*** *1* ***1*** *1011100* $\rightarrow$ *00111* ***1*** *1* ***1*** *1* ***1*** *11100*, where at each step ("$\rightarrow$") the condition (2) "one blocking event equal to/longer than three days and one blocking day separated by a hole" is verified.
A different result is obtained with a similar example: *0010101011100* $\rightarrow$ *0010101* ***1*** *11100* (i.e. one blocking event that is 5 days long). In this case, the vice versa of the condition (2) (i.e. "one blocking day and one blocking event equal to/longer than three days separated by a hole") is verified once.

We write below our algorithm's results for the examples asked by the referee.

00110110100 $\rightarrow$ *0011* ***1*** *11* ***1*** *100* : first condition (1) and then condition (2) are applied;

001110101100 $\rightarrow$ *00111* ***1*** *1* ***1*** *1100* : condition (2) is applied twice;

0011011101110100 $\rightarrow$ *0011* ***1*** *111* ***1*** *111* ***1*** *100* : first condition (1) is applied twice, then condition (2) is applied.

We would like to reiterate that the modifications of the sequence (i.e. the suppression of holes via the conditions (1) and (2)) are very rare: the sequence of labels (4530 elements) is modified 13 times on average, i.e. around 0.3% of the length of the sequence. Moreover, we would like to point out that the condition (2) is verified less than 30% of the times. We added one sentence in the manuscript to stress this aspect: "Overall, the number of holes that are converted into blocking days is very small, about 0.3% of the number of winter days (4530)." In this way, we would like to reassure that the "questionable" cases are few and cannot affect our results.

6. *Fig. 1 caption: Related to my previous comment 15, how about you say in the caption that it is the CRMSD, not RMSD?*

Writing RMSD in the caption is correct. In fact, our current Fig.1 shows the RMSD values computed with $\Delta Z500_{HIST}$, $\Delta Z500_{SSP2}$, and $\Delta Z500_{SSP5}$, while the old Fig.1 (i.e. the one in the manuscript of the first submission) showed the CRMSD values computed with $\Delta Z500_{HIST}$, $\Delta Z500_{SSP2-HIST}$, and $\Delta Z500_{SSP5-HIST}$. We are sorry if we were not clear in our previous reply to comment-15.

7. *Supplement Step A: Related to my previous comment 23, can there be more than one blobs that contain a DG-grid box? How are they treated?*

It rarely happens that DG-grid boxes form different "blobs" and identify different blobs of $\Delta Z500$, like Day 1 in Figure 1. If this happens, since the "scan" of the domain by the algorithm starts from the top-left (or North-West) corner, only the first encountered blob (between 30W and 50E) is retained. On Day 1, only the blob over Scandinavian is retained (last row in Figure 1), however, since the next day, any "blob of DG-grid boxes" (second row) detects the same blob of $\Delta Z500$. We observed that this is the typical situation in those rare cases in which there is more than one "blob of DG-grid boxes".
Overall, most of the times there is only one "blob of DG-grid boxes" (or even only one DG-grid box) per day and, when there are two "blobs of DG-grid boxes", these usually identify the same blob of $\Delta Z500$ (like Day 2 and Day 3 in Figure 1); the case of Day 1 is possible but unlikely.

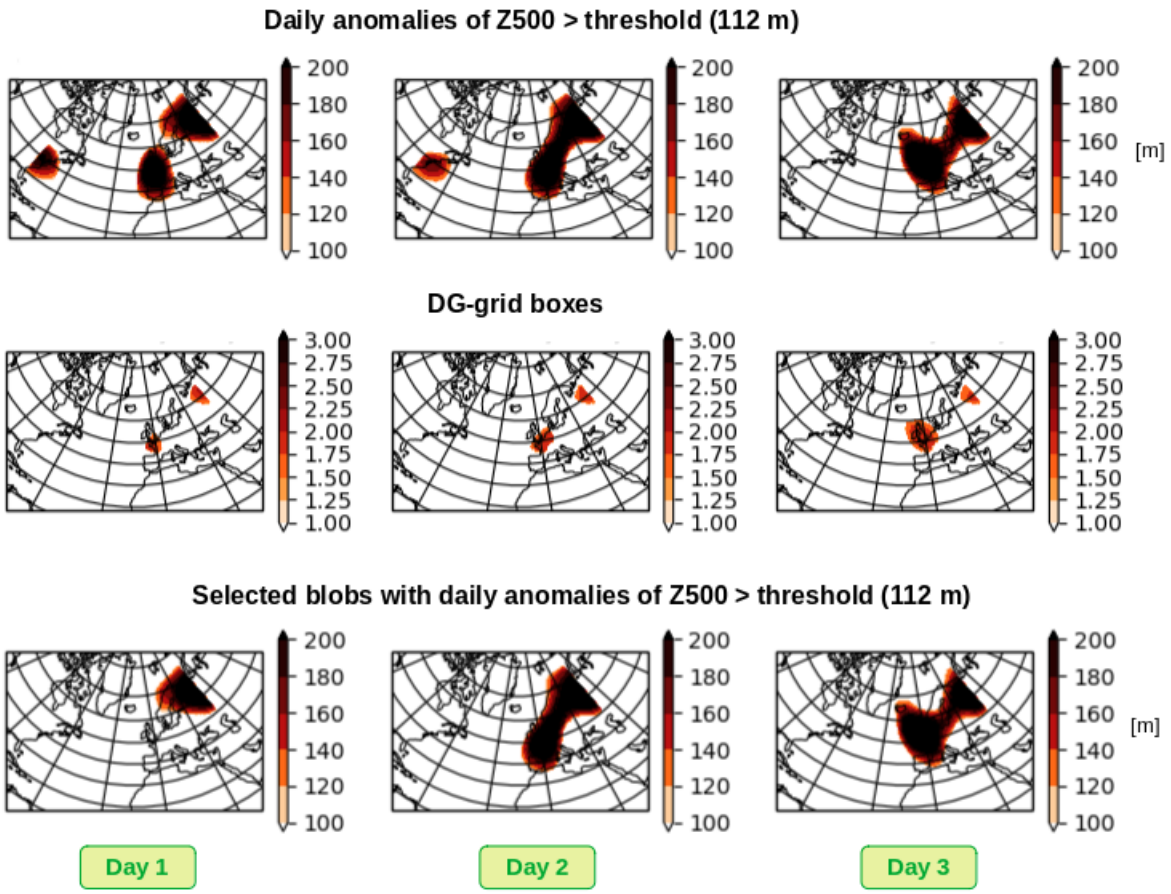
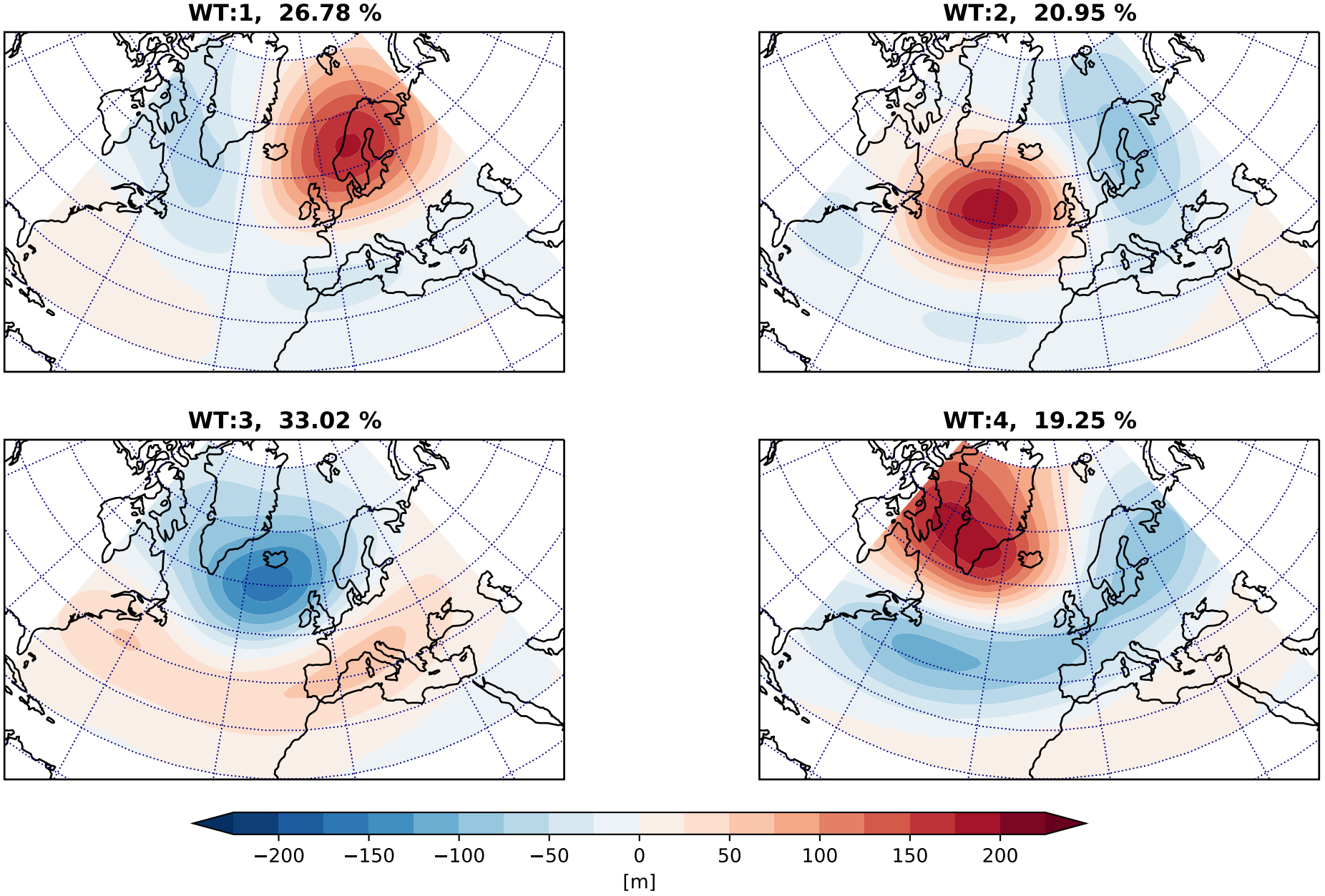
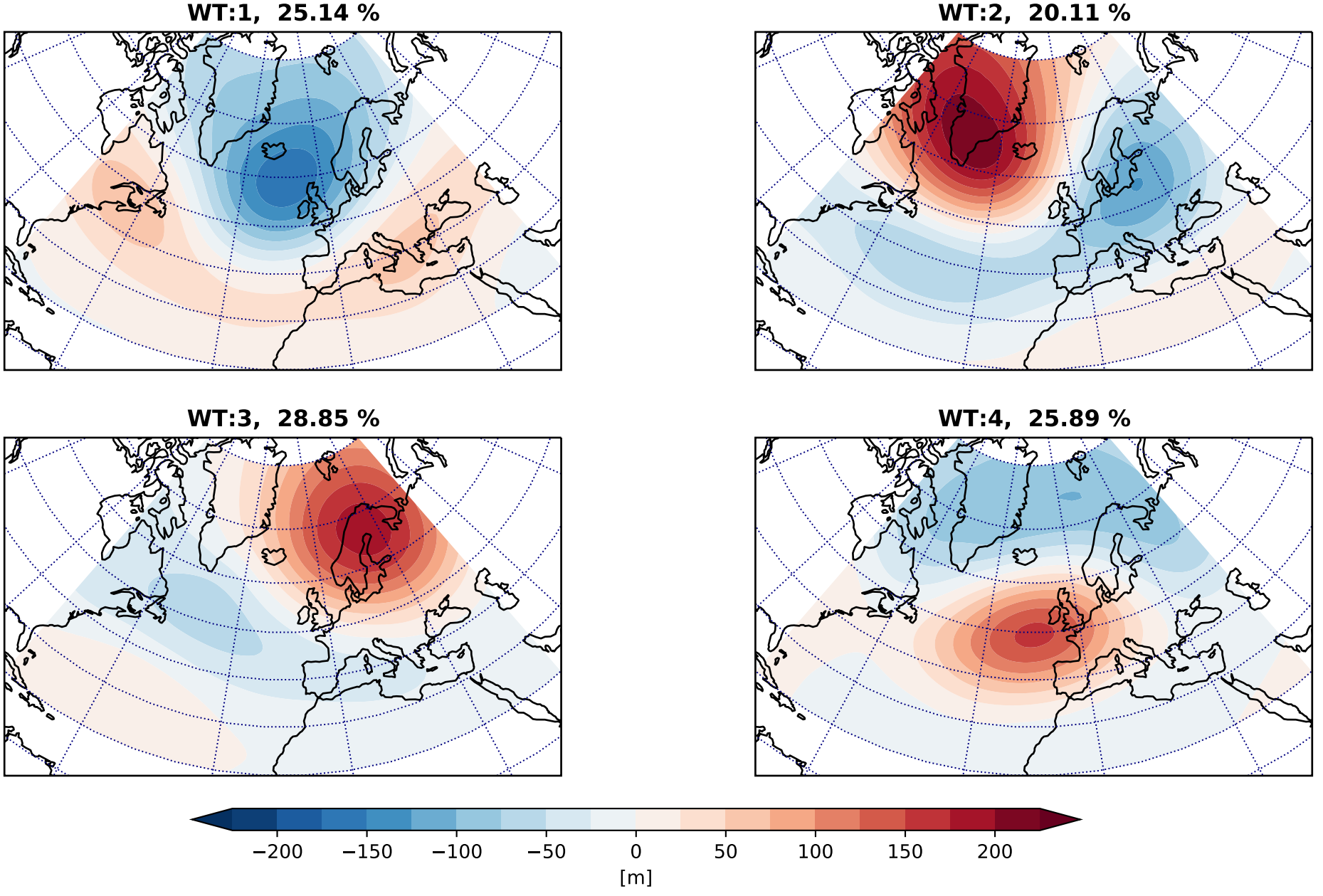


Figure 1: Three days for which the DG-grid boxes are divided in two blobs (see second row). First row: daily $\Delta Z500 > 112$ m; second row: DG-grid boxes; third row: blobs of $\Delta Z500$ selected according to the DG-grid boxes of the second row.

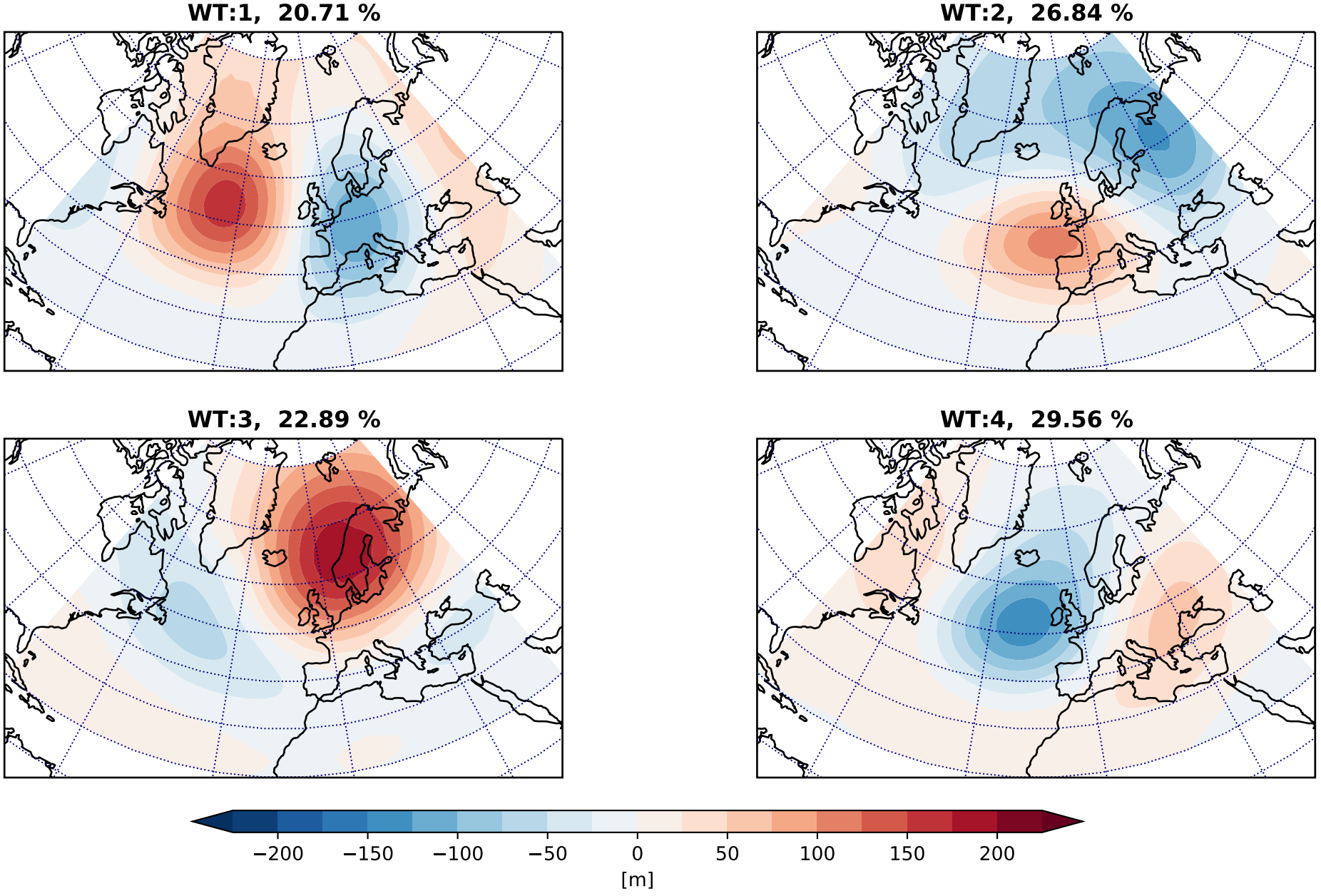
Weather Types (WTs) predicted by ERA5 (hist: 1980-2009)



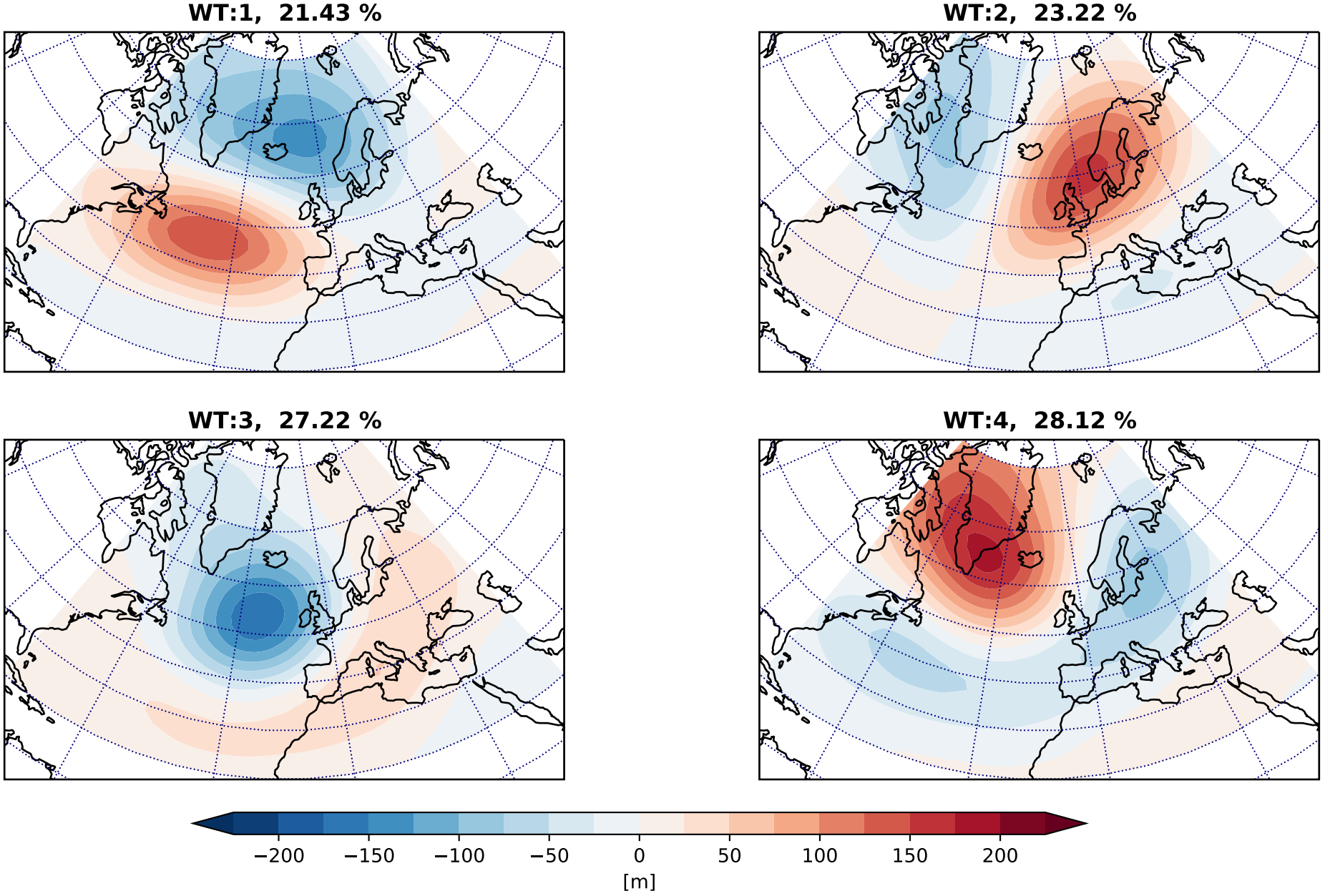
Weather Types (WTs) predicted by BCC (hist: 1980-2009)



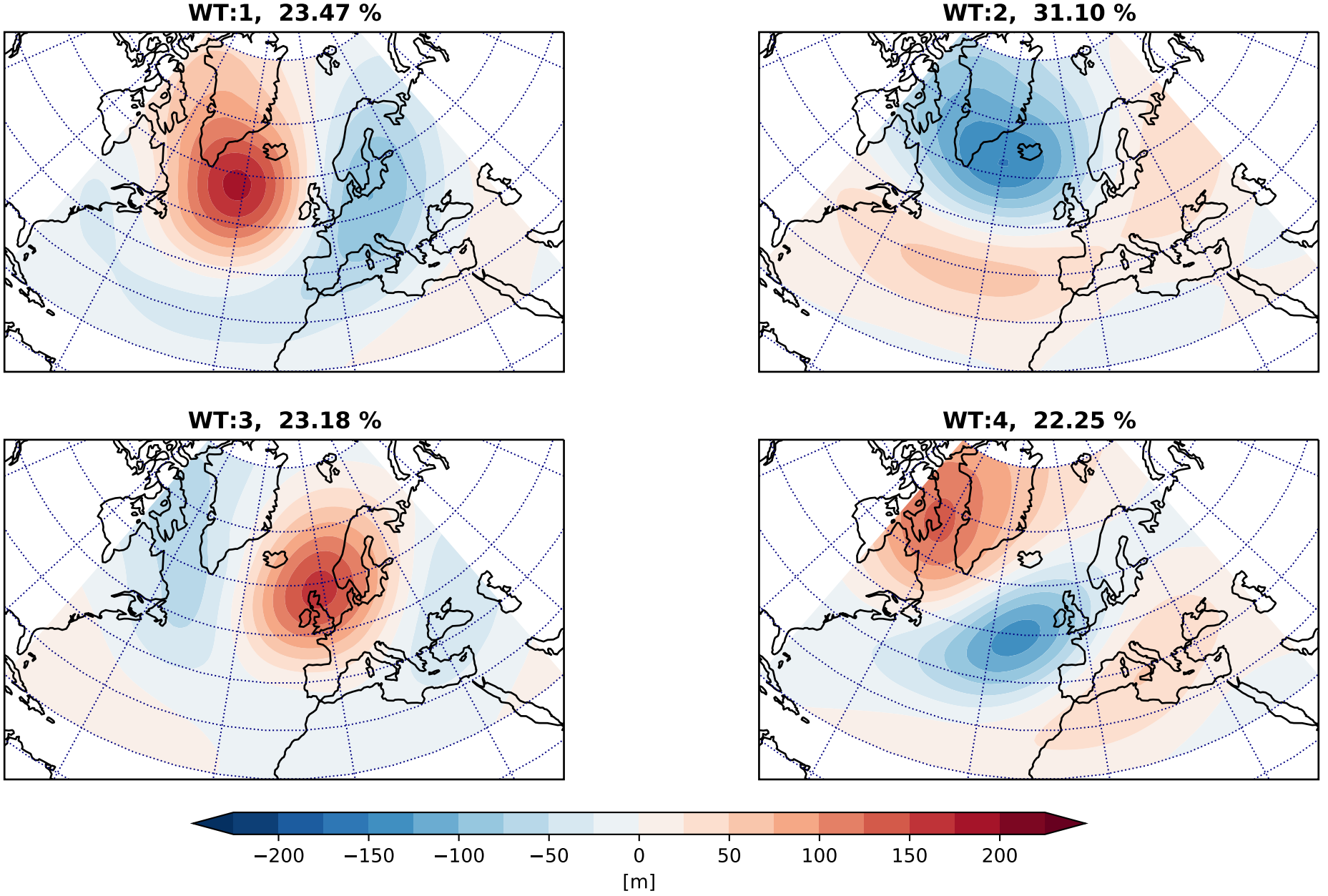
Weather Types (WTs) predicted by FGOALS (hist: 1980-2009)



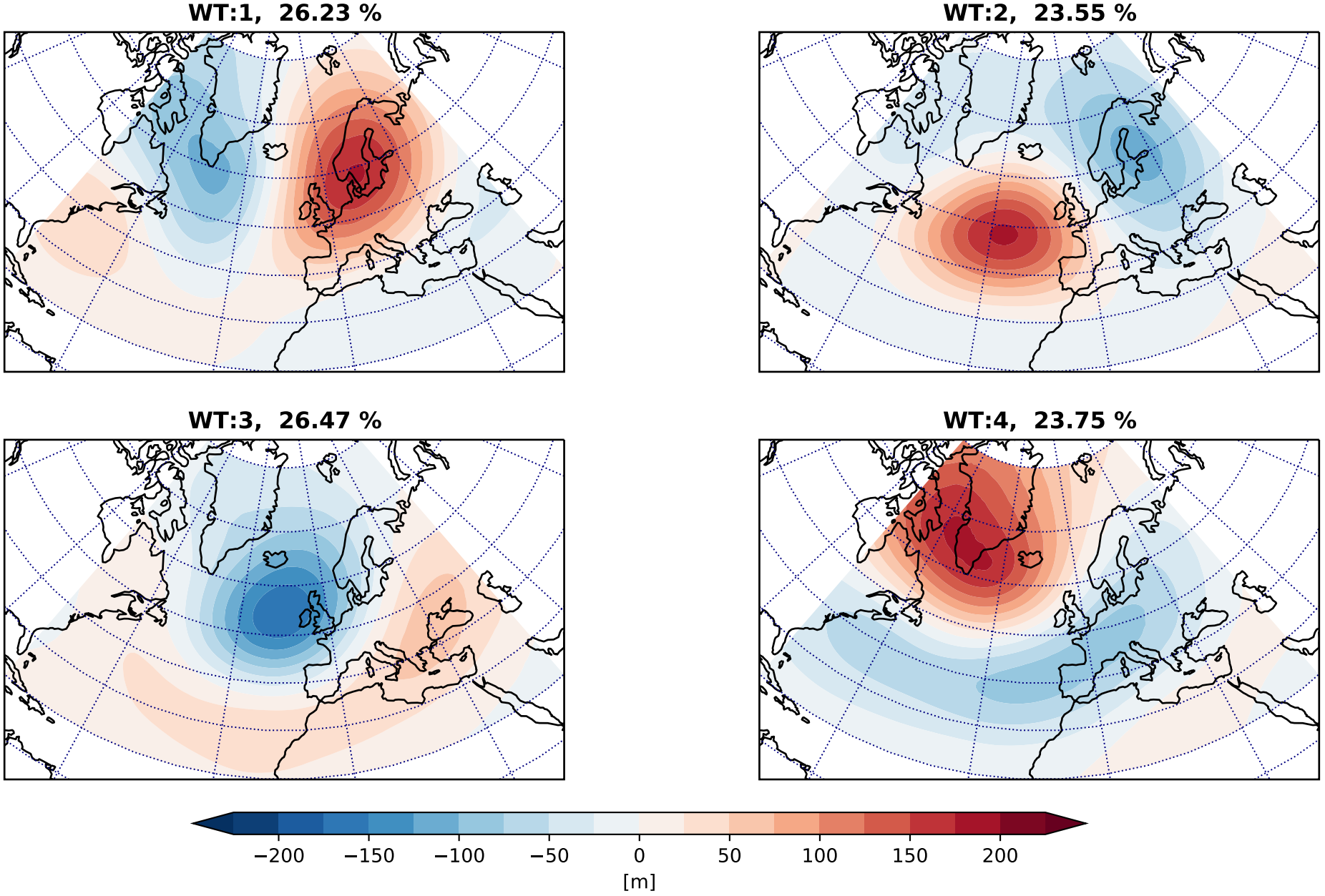
Weather Types (WTs) predicted by GFDL (hist: 1980-2009)



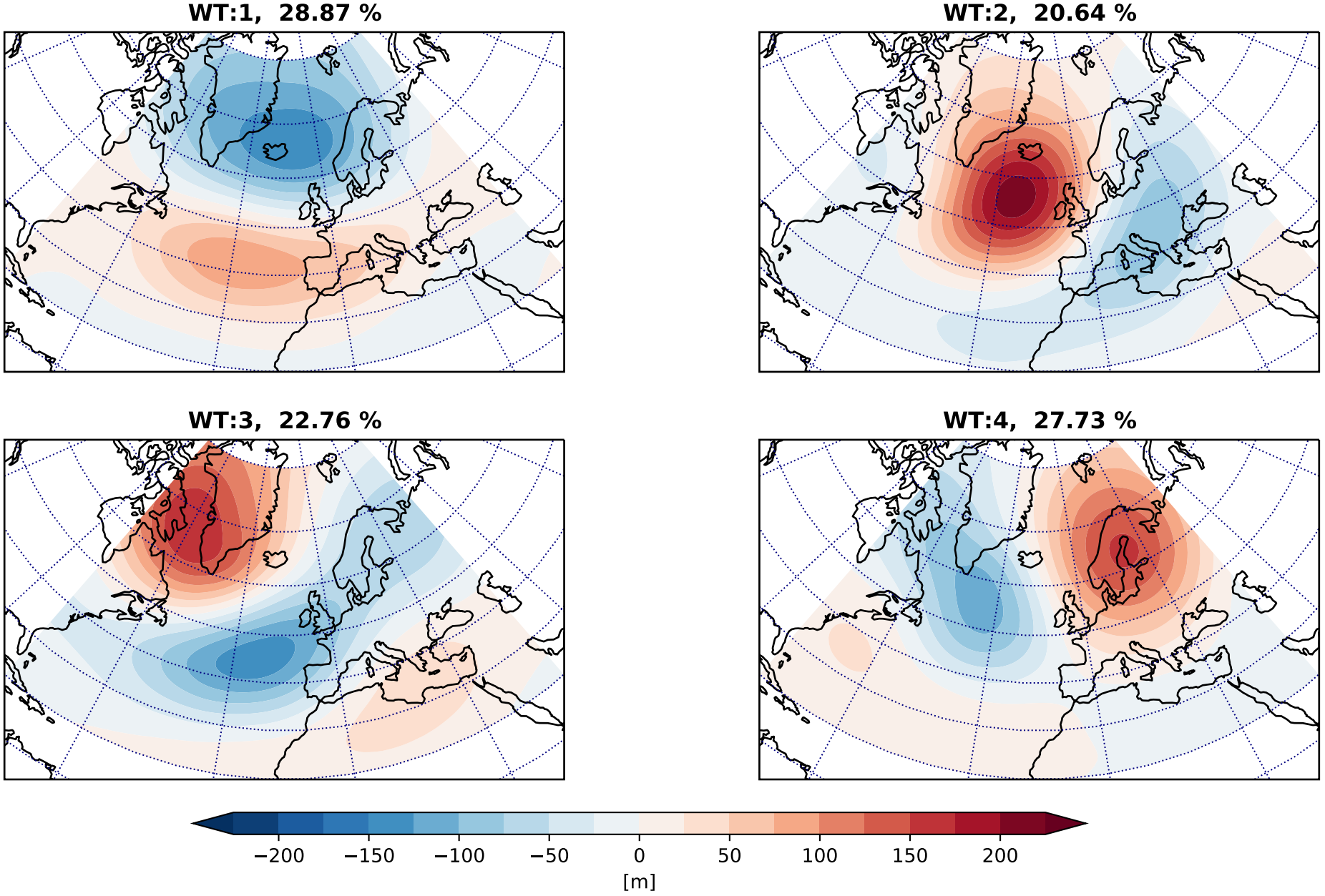
Weather Types (WTs) predicted by INM (hist: 1980-2009)



Weather Types (WTs) predicted by MPI (hist: 1980-2009)



Weather Types (WTs) predicted by MRI (hist: 1980-2009)



Authors' replies to Referee #2

We really thank the referee for the overall comment and useful remarks, which were helpful to improve the manuscript. Below, we provide our replies.

Minor Comments

Line 42, you write: "In fact, WTD allows to explain most of the atmospheric variability . . . " Awkward grammatically. Possible alternatives:
"In fact, WTD can be used to explain most of the atmospheric variability . . . "
"In fact, WTD allows for the explanation most of the atmospheric variability . . . "

In the new version of the manuscript, we used the expression: "In fact, the WTD can be used to explain most of the atmospheric variability . . . "

Line 46: precipitations should be precipitation.

Corrected.

Line 53-55: you write: "For example, it has been shown that the Arctic amplification, which has a strong influence on mid-latitude atmospheric circulation, modulates the frequency and the intensity of blocking events (e.g. Hassanzadeh et al., 2014)." I am sorry that I did not notice this on the first review. I don't think what you write here matches with the intent of the 2014 paper. In fact, that paper was highlighting the need for more study of the possible link between arctic amplification and mid-latitude blocking. Perhaps the editor can add a comment to clarify on this one.

We thank the referee and the editor for pointing out this inaccuracy. This sentence was changed to "For example, the Arctic amplification has been studied in relation to the intensity and frequency of blocking events; although some studies suggest that the Arctic warming yields to an increment of these two quantities (e.g. Francis and Vavrus, 2012), further investigations are necessary to define the Arctic amplification response to blocking (Hassanzadeh et al., 2014; Barnes et al. 2014)."

Line 141: In my first review, I brought up my issue about the use of the term composite. I don't think I explained myself well though. In my view, compositing is an accepted and specific analysis technique used in climate science, it involves calculating the time-average of a phenomena or process identified at multiple time periods. For instance, people often time make composites of sea surface temperature anomalies during the positive phase of ENSO. There has also been a lot of work on cyclone-centered compositing, in which a field such as wind speed is identified for multiple different cyclones and then averaged together to create a composite mean. See for instance, Catto et al. 2010, Fig 3 gives an explanation and Fig 5-7 are examples.
At line 138 you define your own "composite" as the time-average of the delta-Z500 over the blocking days of an individual blocking event. This is time-averaging so that seems ok. But then at line 141 you define a "composite method" that is completely new. For me, this is a bit dangerous, given that there is an existing composite method. I think there will be readers who will be confused about your choice of this terminology. I acknowledge that you made this adjustment to respond to my comment, so I apologize that I did not more clearly explain my

issue. I hope now it is clear and you can choose a different set of terminology so that your new method is not confused with one that already exists. Hopefully find/replace will allow you to do this without much effort.

We thank the referee for the explanation and suggestion. We agree with the arguments of the referee and we changed the name of the method from "composite method" to "WTD method". We think that also this new terminology highlights the difference with the DG method, so it satisfies the comment of Referee #1 in the first report ("*I suggest to change "center method" to "composite method" because "composite" is really the step that differs from the "DG method*"). On the other hand, we preferred to keep the term "composite" with the meaning of temporal mean, as noted by the referee, as this allows us to write the expression "composite of blocking event" to mean one blocking event that is temporally averaged.

Lines 278-285: I don't see the need for Figure 8 right panel. With the revised Figure 2, you have already made the point that the mean state changes affect the detection of the blocks. Here that same point is being made. The text on this panel is also a bit confusing as you use the interpretation of these distributions to discussion a change in blocking size in the future. But that's not what is happening is it? The blocks, the anomaly in the flow, are not changing, as shown with the results in the left panel of Figure 8. If you insist on keeping the right panel of Figure 8, I think more effort needs to be made in the text to clarify that these results represent a change in the mean state rather than a change in the circulation anomalies.

We removed the right panel of Fig. 8; the text has been modified accordingly. We also removed Fig. 6 (right), so we joined Fig. 5 with Fig. 6 (left).
Moreover, we joined the old Fig. 8 (left) with Fig. 7 (left) as Fig. 7 is called in section 4.4.1, while we moved the old Fig. 7 (right) to Fig. 8 as this is cited in section 4.4.2.