Author's Response

M. Schuster

August 25th 2019

Contents

1 Point-by-point response to the reviews

1.1 Response to Comments of Anonymous Referee #1 [R1]

- R1 – comment 1:
  p1, l.20: the first sentences sound as if extra-tropical circulation is important because it may be linked to extreme events. Isn't it important in a more general sense? After all, it is not a paper on extremes. Will be good to discuss the motivation in a broader context
  Response to R1 – comment 1:
  The paper is partly on extremes, as windstorms are identified by the exceedance of the local 98th percentile of the surface wind and blocking is identified by blocked flow for min. 4 consecutive days – these are extreme events by definition. However, the extratropical circulation is indeed important in a more general sense - thank you for the remark. The original text was changed to: "The extra-tropical circulation plays an important role for the redistribution of energy in the atmosphere. The prevailing westerlies and the embedded cyclones and anticyclones determine the weather and climate of the mid-latitudes, assisting in balancing temperature and humidity contrasts between tropical and polar regions. Natural climate variability as well as externally forced climate change determine fluctuations in the circulation and thus i.a. the frequency of extremes such as strong cyclones, intense windstorms or phases of blocked flow. The

consequences of such features include extremes in temperature, precipitation/drought and wind speed, often accompanied by immense damage and harm (e.g. Leckebusch2004, Ulbrich2009, Sillmann2009, Pfahl2012, DeutscheRueck2018)."

- R1 – comment 2:
  The term 'stormtrack' is confusing when used along with the cyclone frequencies – they are sometimes used interchangeably (not in this paper). Though the Methods describe what is meant by the stormtrack, I recommend commenting on the difference early in the manuscript (maybe even in the abstract)
  Response to R1 – comment 2:
  In the abstract p.1,l.3 we inserted the word "different" to emphasize that the stormtrack and cyclones are not the same quantity. The reader can learn about the details of the methodologies in section 2.2.

- R1 – comment 3:
  The same goes to lead years/winters - it is worthwhile explaining which months are considered. I only found this information in Figure captions
  Response to R1 – comment 3:
  The fact that we are analyzing the winter half year (Oct-Mar) is e.g. stated at p.4, l.6: "We will therefore focus on the winter circulation and evaluate averages of the stormtrack and blocking, cyclone and windstorm frequencies from October through March."
  However, to comply with both reviewers' requests for more specific information on the evaluation procedure, the entire paragraph (p.5, l.1ff) was revised to be more precise and now reads:
  "To derive the deterministic skill of the two forecast systems, we focus on the temporal variability and analyze the anomaly correlation for the winters 2-5 (Oct-Mar), following the Decadal Climate Prediction Project (DCPP; Boer2016) protocol. That means that we calculate lead time dependent anomalies of the circulation measures. This is a simple and robust approach to account for a possible lead time dependent mean bias, i.e. drift. Thus, for each of the initialization experiments (1978, 1979, ...) the ensemble average (5 members) of the temporal mean of the 4 contained lead winters is calculated per grid point. This forms a new ensemble mean time series of the lead winters 2-5. This time series serves to calculate the climatology (temporal mean) as well as the respective anomaly time series. The time series of those anomalies of the hindcasts is then correlated (Pearson) to the time series of anomalies of the reanalysis. In decadal prediction studies, this procedure is usually repeated for each lead time, e.g. lead year 1, lead year 2-5, lead year 6-9 - it is therefore referred to as lead time dependent anomaly correlation. In our study we only show results for one lead time: lead winters 2-5. The initialization of the hindcasts takes place in October, this means the first full winter that we analyze is the second winter, i.e. the months 12-17 (Oct-Mar) after initialization.

This evaluation procedure is part of the decadal climate prediction evaluation software that was designed within the MiKlip project (Illing2014) and is applied for this study. This OpenSource evaluation software follows the evaluation framework of (Goddard2013) which led to the DCPP requirements. “

- R1 – comment 4:
  p2, l28: comment on what parametric bias adjustment approach is.
  Response to R1 – comment 4:
  The wording parametric or non-parametric corresponds to the way how to adjust lead time dependent bias (drift). On the one hand, it is feasible to assume a lead time dependent bias and to fit a curve. Kruschke et al. (2016) called this approach parametric. On the other hand, DCPP recommends to calculate lead time dependent anomalies for each lead year separately (Boer et al., 2016). This is a non-parametric approach. We used the DCPP recommendation in our manuscript.

- R1 – comment 5:
  p10,l1: The word ‘shift’ often implies change in time, consider revising
  Response to R1 – comment 5:
  We think that the structure of the sentence makes it clear, that a spatial shift in LR compared to the reanalysis is meant.

- R1 – comment 6:
  p10,l13: I would be more precise here and stick to the words used in the Methodology, i.e. ‘open’ and ‘closed’. Otherwise, you need to clarify what you mean by weak/strong cyclones.
  Response to R1 – comment 6:
  The following sentence was added to p.6, l.20: “Only cyclones that live for more than 24 hours and reach a Laplace larger than 0.7hPa/(degree latitude)^2 and have closed isobars at least once during their lifetime are selected for evaluation.”
  p.10 l.13 was changed to:” However, it should be highlighted that the cyclone tracking algorithm also detects weak and moderate cyclones.” This means, that a cyclone can be part of the evaluation, which lives a couple of days but is generally weak in terms of its Laplacian of the pressure ($<$ 0.7hPa/(degree latitude)^2), but it reached the intensity criterion for exactly one time step and therefore was included in the evaluation.

- R1 – comment 7:
  p.10,l14: I would like to see a figure confirming that positive bias is due to the weak and/or short leaved cyclones. p.10,l.15-16: How do you explain then negative windstorm vs positive stormtrack anomaly over the Atlantic?
  Response to R1 – comment 7:
  To answer this question we selected and evaluated cyclones that, at any time during their lifetime, pass through the central North Atlantic (50°-10°W,40°-60°N) - the region where the bias in Fig. 3a is strongest. This

analysis is performed for individual cyclone tracks of all initialization experiments between 1960-2012, all 9 forecast winters and all 5 members, for LR and HR respectively.

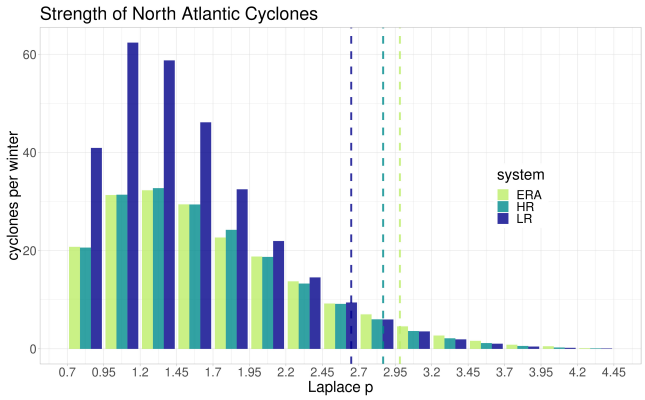


Figure 1: Intensity histogram (max. along track Laplace of pressure) of cyclones that passed the central North Atlantic (50°-10°W,40°-60°N) once during their lifetime in the different decadal forecast systems (LR, HR) and the reanalysis.

The intensity histogram (review response Fig. 1) of cyclones that pass the central North Atlantic shows that weak cyclones are more numerous in LR than in HR or ERA-Interim. Although LR overestimates the frequency of weak cyclones in that region, the frequency of the strong cyclones, in that case the strongest 5% of cyclones, i.e. bars to the right of the dashed line, is reproduced quite well in LR. This threshold (dashed lines) in HR (2.87 hPa/(deg.lat.)^2) is closer to ERA-Interim (2.98 hPa/(deg.lat.)^2) than LR (2.65 hPa/(deg.lat.)^2) is to ERA-Interim - but mainly due to the generally larger number of events in LR. Overall, the shape of the intensity distribution for cyclones passing that region is much more similar, and actually almost identical, between HR and ERA-Interim, than between LR and ERA-Interim.

The lifetime histogram (review response Fig. 2) of cyclones that pass the North Atlantic region also shows that short-lived cyclones in this region are more frequent in LR than in HR or ERA-Interim. Again the shape of the distribution matches very well between ERA-Interim and HR.

Regarding the second part of the comment: The stormtrack is calculated from the variance of the geopotential height in the synoptic band. Within this quantity, there are many systems included which do not produce windstorms, as this is the variability of all geopotential values (high pressure as well as low pressure systems, and strong ones as well as weak ones). Only

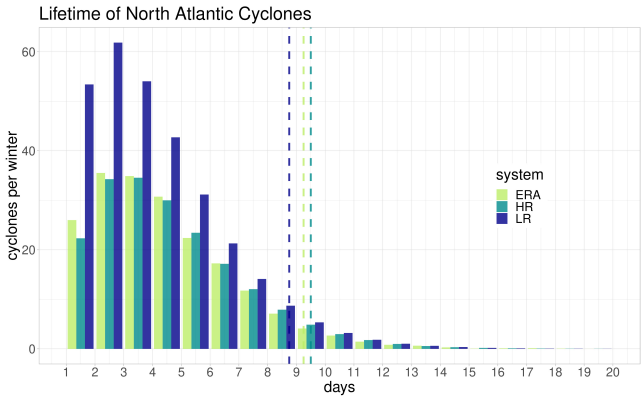


Figure 2: Lifetime histogram (length of track) of cyclones that passed the central North Atlantic (50°-10°W,40°-60°N) once during their lifetime in the different decadal forecast systems (LR, HR) and the reanalysis.

strong low pressure systems can be related to windstorms. It cannot be expected that the signals of total variance of geopotential height (storm-track) are identical to those of the windstorms since the distribution of strong and weak cyclones changes differently as discussed above.

- R1 – comment 8:
  p10, l.31: I can see a discussion on negative correlations further in this section (e.g.p13, l7)
  Response to R1 – comment 8:
  As we lined out in the paper, it is not desirable to have a deterministic prediction model to continuously predict the opposite of the observed quantity. We therefore stick to the opinion that negative correlations should not be considered skillful and will therefore not discuss them in detail. If anything, then the message of our paper is that the amount of negative correlations is reduced in HR. This is covered by the discussion of positive differences between HR and LR (Fig. 4e, 4f, 5e, 5f of the originally submitted manuscript).

- R1 – comment 9:
  p12,l6: I How about a strong reduction of skill over Northern Canada and the Barents Sea
  Response to R1 – comment 9:
  The following sentence was added to p.12, l.6: "However, there is also an area of a significant reduction of the anomaly correlation for the stormtrack over Northern Canada and the Baffin Bay."

- R1 – comment 10:
  p.12, l3: 'significant skill improvement' - the authors probably mean that HR model shows statistically significant correlation with ERA-Interim at

some points. In my opinion, though, this statement makes an impression that skills of model prediction have become really good (so say at least 'statistically significant skill improvement' or rephrase). More important, the prediction skills, as shown in the paper, are remarkably low for most part of the region, but this message is not conveyed by the paper - will be good to see more discussion on that.

Response to R1 – comment 10:

Referring to Fig. 4.e in this line (p.12 l.3), we are not discussing the skill of HR compared to the reanalysis but rather the change in skill from LR to HR. Thus, we indeed mean that the change from LR to HR shows an improvement, e.g. from low or (significant) negative correlations (no skill) in LR to significant positive correlations (skill) in HR, i.e. an improvement in skill or one could also say a statistically significant improvement in anomaly correlation.

Please note that deterministic decadal prediction skill in terms of anomaly correlation is generally low for model variables other then surface temperature (compare skill of precipitation in Kadow et al., 2016 or cyclone frequency in Kadow et al. 2017 to skill of surface temperature in Pohlmann et al., 2013).

We changed the original wording to: "...statistically significant skill improvement..." and added values of correlation in the text to put results into perspective.

- R1 – comment 11:

  Figures 4-5: In line with the previous comment, it will be interesting to calculate the percentage of area that is significantly (positively?) correlated with ERA-Interim. This number can be added to each subplot.

  Response to R1 – comment 11:

  As the climatologies of the circulation quantities show, the values of the anomaly-correlation are not equally important anywhere in the displayed domain. A significant positive correlation e.g. along the maximum blocking frequency is more relevant than a significant positive correlation over a region with very low frequencies. Therefore, we think the suggested percentage of grid points with significant positive correlation would be misleading, as it would weigh "irrelevant" regions equally to "relevant" regions. We produced the figure (review response Fig. 3) to answer the reviewer's question but will not show it in the paper for the stated reasons. The figure shows that for all circulation quantities the number of grid points (in the North Atlantic domain) with significant negative anomaly correlation is reduced in the higher resolution system and the number of grid points with significant positive anomaly correlation is increased in the higher resolution system - supporting the theory of improved physical processes throughout the region.

  We understand the referee's main point with this comment is, similar to R1 comment 10, that we emphasize the positive effects of the resolution

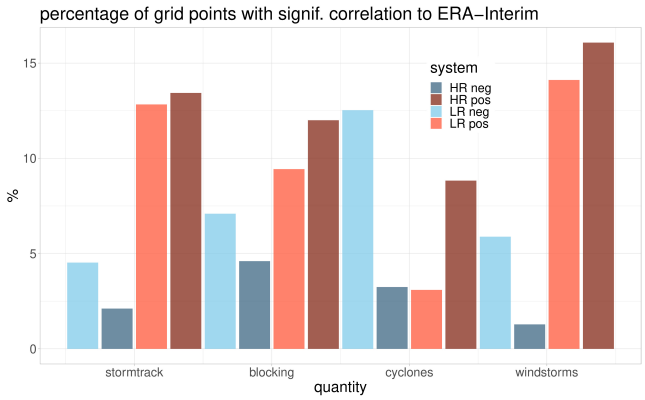


Figure 3: Percentage of grid points within the North Atlantic domain of the different systems (LR, HR) that show a statistically significant correlation to ERA-Interim.

and the reader could think we suggest that HR is the perfect decadal prediction model. This is however not our intention. We rephrased respective sentences.

- R1 – comment 12:
  Discussion and Conclusions: this section is too long, consider shortening. Parts of the discussion may be moved to the Results. The last sentence of the article is not clear, please revise.
  Response to R1 – comment 12:
  We decided to separate the discussion from the conclusion - so the conclusion is a lot shorter now. With respect to the second reviewer's comment we also changed the discussion to be more critical and related our findings to other studies' results.

- R1 – comment 13:
  Fig4: significant at what level
  Response to R1 – comment 13:
  We added "(95% significance level)" to p.5 l.10.

- R1 – comment 14:
  p2,l.11: remove comma before dash. p2, l18-19: put references in brackets p2, l29: did you mean more skilfull ? Skilful is misspelled. p5, l11: 1000 time - should '1000 time steps' be better? p10, l6: should read 'these results'
  Response to R1 – comment 14:
  We replaced the dash with a comma, to separate the two independent but related sentences.
  Thank you for noticing, brackets were inserted.
  No, we do mean simply skillful.

We are using American English throughout the paper, the spelling is correct: skillful.
We use the bootstrap method, to estimate the distribution of a population by resampling the dataset with replacement. This is repeated 1000 times. The suggested term "1000 time steps" is not applicable.
The word "results" is used as a verb here. The suggested use of the word "results" as a noun would change the meaning of the sentence and leave it incomplete.

1.2 Response to Comments of Anonymous Referee #2 [R2]

- R2 - comment 1:
  The applied methods are often not clear. The use of an "evaluation software" is mentioned (P5L3). What does it actually do? When is the ensemble mean calculated, e.g. are the shown correlation maps means of correlations or correlations between ensemble mean and reference. Please provide clarification and add the applied calculation methods/equations. Could be as appendix/supplement.
  Response to R2 - comment 1:
  The evaluation software, as described in p.7, l.6-11, comprises the different post-processing routines to derive the stormtrack and the three different frequencies from the direct model output and it also comprises a routine for the skill (anomaly correlation) analysis. This evaluation software named "freva" was designed within the MiKlip project and used as Central Evaluation System by research groups within this project. Based on standardized model output, the "freva"-user can apply different evaluation or post-processing methodologies in an easy and reproducible way. What these single post-processing routines - or plugins as they are also called - do, is described in Section 2.2. This means, from the direct model output of the hindcasts, first the four circulation metrics and winterly averages of their statistics are calculated for the reanalysis and the hindcasts. Afterwards, lead time dependent anomalies and the anomaly correlation are calculated as follows: For each of the initialization experiments (1978, 1979, ...) the ensemble average (5 members) of the temporal mean of the 4 contained lead winters is calculated per grid point. This forms a new ensemble mean time series of the lead winters 2-5. This time series serves to calculate the climatology (temporal mean) and to calculate the respective anomaly time series. The time series of those anomalies of the hindcasts is then correlated (Pearson) to the time series of anomalies of the reanalysis. In decadal prediction studies, this procedure is usually repeated for each lead time, thus lead year 1, lead year 2-5, lead year 6-9 - it is therefore referred to as lead time dependent anomaly correlation. In our study we only show results for one lead time: lead winters 2-5.
  Hence, the correlation maps in Fig. 4 and Fig. 5 show correlations between the ensemble mean and the reference.

We implemented this description to the manuscript text - see R1 comment 3.

- R2 - comment 2:
  The study suggests a direct relation between the mean bias of the ensemble mean and the anomaly correlation of the ensemble mean to the reference for one and the same diagnostic. The correlation is insensitive to the mean bias on grid cell level, hence anomaly. It appears large parts of the result section and conclusions are based on the assumption that a reduction of mean error/bias leads to higher anomaly correlations for the same analyzed quantity. This has to be revised substantially.
  Response to R2 - comment 2:
  Thank you for the remark. It was not our intention to suggest a direct relation between bias and the anomaly correlation. Rather, the independent, but locally coincident, improvement of both, the bias and of the anomaly correlation, for the same quantity points towards an improvement of the physical processes in the HR model. We assumed we had already chosen our wording carefully. We revised and clarified the respective paragraphs.

- R2 - comment 3:
  The hindcasts are presumably not post processed, e.g. corrected for time-varying bias, trend-adjusted, etc? Please clarify and state why this might be not necessary. Why is the approach of correcting biases of this study different to Kruschke et al?
  Response to R2 - comment3:
  In the third paragraph of Sec 2.1, we give information about how data is post-processed and analyzed. This is apparently not clear enough. Thank you for the comment.
  We analyzed the frequencies of the circulation metrics, i.e. values for each lead winter, respectively, following the DCPP recommendation. That means that we calculated lead time dependent anomalies of those frequencies (see R2 comment 1 and R1 comment 3). This is a simple and robust approach to account for a possible lead time dependent mean bias, i.e. drift (DCPP recommendation, Boer et al., 2016).
  There exist miscellaneous more sophisticated approaches for the post-processing of decadal predictions (Kharin et al., 2012, Kruschke et al., 2016, Pasternack et al., 2018). In our study we wanted to point out the effect of the model resolution on the forecast skill of the circulation measures and therefore, we intentionally did not compare the LR model including a complex post-processing approach with the HR model including a complex post-processing approach.

- R2 - comment 4:
  Spatial resolution has been discussed to be a serious limiting factor to correctly reproduce climate mean state and variability in the context of decadal prediction (e.g. Hewitt et al. BAMS 2017, Smith et al. QJRMS 2016). This should be mentioned more prominently in the motivation and

put in context of this study in the discussion. There are numerous studies about the effect of resolution in climate models in general including the effect on North Atlantic circulation measures (e.g. Davini et al. J ADV MODEL EARTH SY, 2017). How do they compare to this study?
Response to R2 - comment4:
We dedicated an entire paragraph of the introduction (p.3, l.1ff) to this topic. We discussed the limiting factor 'resolution of climate models' and its effects on the representation of the ocean surface state and in particular on the representation of the North Atlantic atmospheric circulation, and we cited a multitude of studies dealing with this topic. Also, we have discussed state of the art results from studies in which higher resolved decadal hindcast sets were analyzed. Nevertheless, we added some of the suggested papers to our citation list, where appropriate.
Added before p.3, l.1:
"It is well known that a coarse spatial resolution of global coupled climate models hinders the proper representation of sub-synoptic scale systems, and thus the climate mean state and variability."
Added to p.3 l.10:
"Similar effects for the blocking frequency bias are found in an atmosphere only model by Davini et al. (2017)."
Added Hewitt (2016) to paper listing:
"It has been found in many studies, that the atmospheric dynamics benefit not only from a coupling of the atmosphere and ocean but also from an increased model resolution (Shaffrey, 2009; Jung 2012; Dawson, 2013; Hewitt 2016)."

- R2 - comment 5:
The difference in mean bias for LR and HR in cyclone frequency is striking and given the very small differences in stormtrack activity somewhat unexpected, e.g. at 30W, 50N (Fig 2a vs Fig 3a). The result is apparently similar to Kruschke et al 2014. Kruschke et al compare uninitialized experiments in LR to NOAA's 20th Century Reanalysis. In their study the mean bias is up to 25 systems per winter over the North Atlantic and they mention a possible underestimation of cyclone frequency of the reanalysis. This seems at odds with what is shown here: A mean bias of up to 80 systems and more per winter in comparison to a different reanalysis product. Please discuss this. Is it possible to estimate how much is due to the applied tracking method? One suggestion could be to interpolate the HR hindcast to the lower resolution and repeat the analysis. Will that change the results? This could be done for a single member and put as appendix. It is mentioned that LR overestimates weak and moderate systems. Why?
Response to R2 - comment 5:
Regarding your suggestion to interpolate HR to the lower resolution: Usually the experience is, that the finer the resolution of the model, the more accurate the description of the pressure field and the more cyclones can

be detected by the algorithm. So, if we interpolated HR to the lower resolution, we would not expect to see an LR-like positive bias - rather the opposite is the case, we would expect even less cyclones in the interpolated HR hindcasts. This would not be helpful to explain the strong positive cyclone bias. We therefore decided not to follow this suggestion. We understand, that your question points towards an explanation for the strong positive cyclone frequency bias in LR. This question is already partly answered in the response to R1 comment 7 (positive cyclone frequency bias is produced by weak and short-lived cyclones) and is complemented by the next few paragraphs.
The cyclone identification and tracking method applied in our study is identical to the one used in Kruschke et al. (2014) - the methodology originally designed by Murray and Simmonds (1991) - so the differences in cyclone frequency biases in the two studies cannot be derived from a different methodology. The same holds for the computation of the frequencies, in both cases the frequency was derived from cyclone counts within a distance of 1000 km around a grid point. The differences can however be explained by the different datasets used. In Kruschke et al. (2014) the bias of the un-initialized LR runs (of an older MPI model version) relative to 20CR is shown. In our study the bias of the initialized LR runs (of the current MPI model version) relative to ERA-Interim is shown - so the MPI model version differs, the initialization differs and the reanalysis dataset differs. We performed a few studies in the attempt to isolate the different effects.

Effect of the new model version
To test the influence of the model development on the bias, we analyzed the cyclone frequencies in the un-initialized MPI-ESM runs used and shown in Kruschke et al. (2014) and in the respective un-initialized runs of the MPI-ESM model used in our study - please note, that we never showed results from the un-initialized runs, but only from the initialized runs in our paper.

This model development from the system used in Kruschke et al. (2014) termed 'Baseline1' (B1) to the current MPI-ESM system termed 'Preoperational' (Preop) slightly reduces the winterly cyclone counts over the North Atlantic (review response Fig. 4). The effect is negligible (-4 cyclones per winter) compared to the strong bias we see in the initialized Preop-LR, and is of opposite sign. Thus, the model development alone cannot explain the strong North Atlantic cyclone frequency bias.

Effect of the initialization
The comparison between the initialized Preop-LR runs used and shown in our study and the respective un-initialized runs of the Preop-LR system however shows a very strong increase in North Atlantic cyclone frequencies

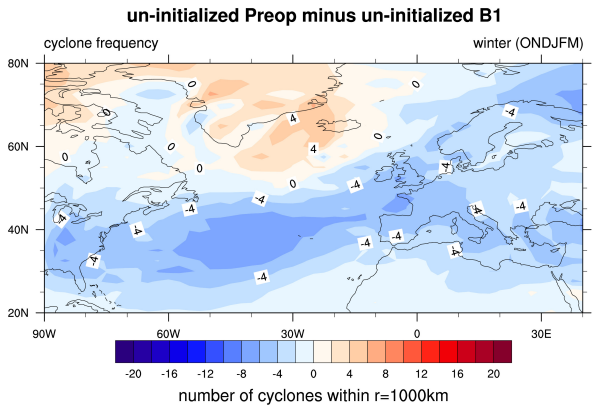


Figure 4: Effect of the model development - Difference of the average cyclone frequency between the un-initialized Preop-LR and un-initialized B1-LR simulations.

(+100 cyclones per winter; review response Fig. 5). This indicates that the majority of the bias seen in Fig. 3a of our study can be explained by the initialization of the Preop-LR system.
Actually, this initialization effect is also inherent in the older B1 system (review response Fig. 6), between the un-initialized runs used and shown in Kruschke et al. (2014) and the respective initialized runs of the same system also used in Kruschke et al. (2014) - but they only showed the bias for the un-initialized runs in their paper.
Given the fact that the initialization technique in Preop-LR and Preop-HR is identical, but only LR exhibits the strong cyclone frequency bias, it appears to be an unfavorable interaction, between the LR system and the initialization, which triggers this bias. In the following we explored what this interaction might entail.

Taking a closer look into the initialized LR system, we find a negative sea-level-pressure bias over the central North Atlantic. This is shown in the review response Fig. 7 (left) for the initialized simulations used in our study; and for the un-initialized simulations of the same model version in Müller et al. (2018, their Fig. 7c). The systematically too low pressure over the central North Atlantic seems to affect existing flow disturbances, i.e. weak/open cyclones, over the central North Atlantic, by strengthening them and artificially extending their lifetime just enough to meet the algorithm's thresholds, so that a strong bias in the average cyclone frequency becomes visible. As shown in the intensity and lifetime

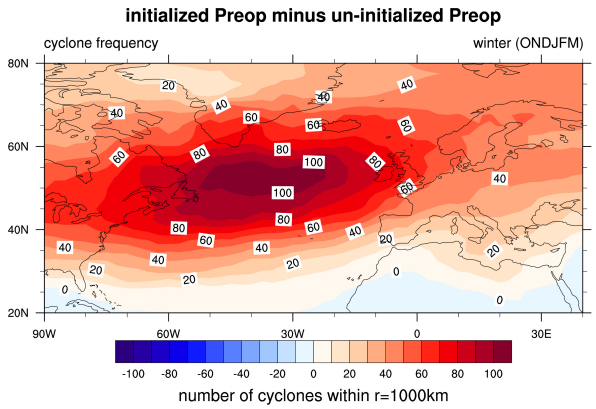


Figure 5: Effect of the initialization in Preop-LR - Difference of the average cyclone frequency between the initialized Preop-LR and un-initialized Preop-LR simulations.

histograms, in response to R1 comment 7, this bias can be attributed to weak and short-lived cyclones. Obviously this pressure bias in LR acts to produce artificial cyclones.
Although a negative pressure bias is still visible in HR (review response Fig. 7, right) but shifted to Newfoundland, we do not see a likewise strong bias in the cyclone frequencies there. The negative pressure bias in the cyclogenesis area (Newfoundland) seems not to be as critical. We conclude, that the negative pressure bias in the two hindcast systems is more relevant for existing disturbances (strengthening those to become weak and moderate cyclones over the North Atlantic in LR) than for the genesis of cyclones (over Newfoundland in HR).

Effect of the reanalysis

To round off the picture, we compared the cyclone track density biases of the initialized and un-initialized MPI systems relative to different reanalysis datasets. The plots in review response Fig. 8 are for the B1 system, but they look essentially identical for the Preop-LR system. The bottom, left figure corresponds to the bias seen in Kruschke et al (2014) - a bias of 20-30 cyclones over the Eastern North Atlantic and Europe for the un-initialized system relative to 20CR. If they had used ERA-Interim instead of 20CR the top, left figure would have appeared - a general underestimation of the un-initialized B1 system over the Northern North Atlantic of -20 to - 40 cyclones. The comparison between the left and right column illustrates again the initialization effect. The top, right figure is equivalent

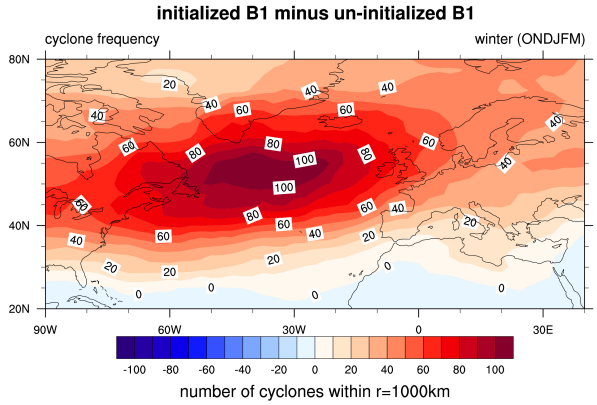


Figure 6: Effect of the initialization in B1-LR - Difference of the average cyclone frequency between the initialized B1-LR and un-initialized B1-LR simulations.

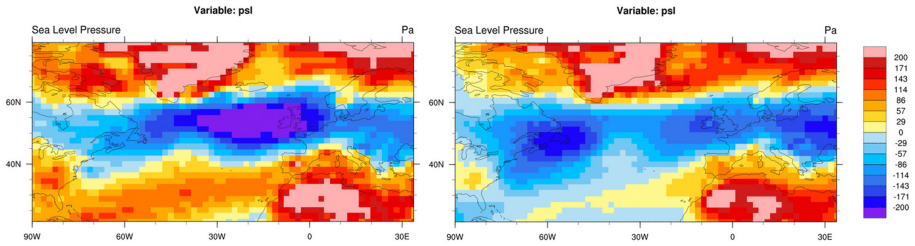


Figure 7: Mean Sea-Level Pressure bias relative to ERA-Interim - left: in the Preop-LR system; right: in the Preop-HR system.

to the bias shown in our study - a bias of +80 cyclones over the central North Atlantic in the initialized system relative to ERA-Interim.

- R2 - comment 6:
  The ensemble spread is unfortunately not used or shown for any of the analyses. How is the spread different between LR and HR? Is the reanalysis within the spread?
  Response to R2 - comment 6:
  Instead of checking whether the reanalysis is within the spread, we follow the CMIP or DCPP suggestion to compare the ensemble spread with mean squared error of the model compared to the reanalysis - to see if the spread is an adequate representation of the uncertainty. The spread is equally strong in LR and HR and close to the MSE (applying the Log.

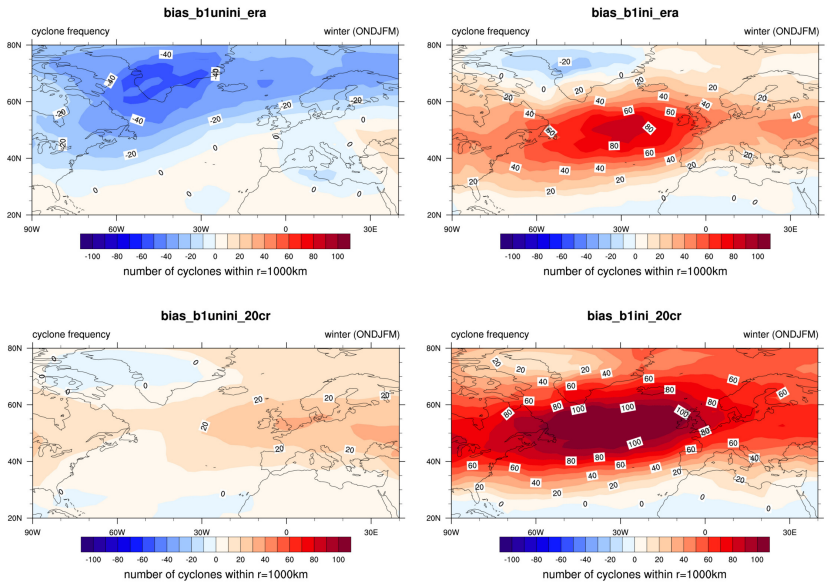


Figure 8: Cyclone frequency bias in the different simulations relative to different reanalyses - top: relative to ERA-Interim; bottom: relative to 20CR; left: uninitialized simul.; right: initialized simul.

Ensemble Spread Score) for each of the respective quantities (stormtrack, blocking frequencies and windstorm frequencies - not shown). For those quantities it is not necessary to show the plots. Only for the cyclone frequencies (review response Fig. 9), the spread in LR is larger than in HR over the North Atlantic, i.e. where the bias is high, and over Eastern Europe. This means that additionally to the average cyclone bias, created by the North Atlantic pressure bias and the initialization (as discussed in response to R2 comment 5), the members produce largely varying numbers of cyclones per winter. This result is in agreement with the bias in weak cyclones as shown in R1 in comment 7. Still, the ensemble spread in LR is not overwhelmingly high. We added two sentences to the manuscript.

- R2 - comment 7:
  When analyzing absolute numbers (here for blocking, cyclones and windstorms) ties have to be considered in the correlation calculation, i.e. seasons with the same number of events. Presumably ties are not taken into account as the manuscript does not mention it. 2 possible solutions: i) mask regions with a large number of ties ii) use a different correlation coefficient, e.g. Kendall's Tau B. Otherwise the correlation value could be misleading and statistical significance becomes meaningless, especially in

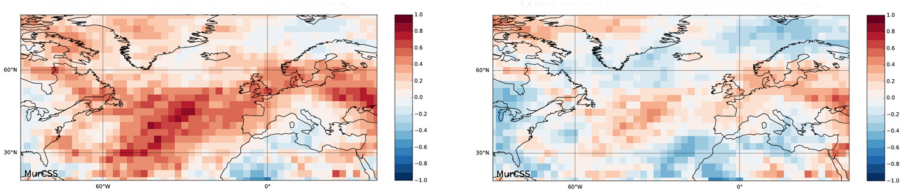


Figure 9: Spread vs. MSE (Logarithmic Ensemble Spread Score - LESS) for the cyclone frequency - left: in the Preop-LR system; right: in the Preop-HR system

regions with few events per season. There is a significant negative correlation in windstorm frequency in LR over Eastern Canada and a significant positive correlation in HR over the same region. This could be an example of too many ties.
Response to R2 - comment 7:
In the manuscript there is a lag in the explanation of how the time series are preprocessed before correlations are calculated. Thank you for this feedback! We added a more detailed description to Sec. 2.1 where this is explained. It includes the information of anomaly, ensemble mean and running mean computations. Due to this type of preprocessing we decided to use the Pearson correlation coefficient instead of rank correlations - the latter would have been affected by ties. Nevertheless, we analyzed the number of ties and found, that due to the ensemble mean and running mean, there are almost no ties in the hindcasts, and only few ties in the reanalysis data.
Significance of the correlation is calculated by means of a bootstrapping, resampling the time series with replacement. Ties (only few cases as mentioned) are also used for the bootstrapping which leaves significance still meaningful.

- R2 - comment 8:
  Related to the above point: Cyclone frequency is apparently masked in regions with high orography. This can be seen in Fig 5. Why is there no mask in Fig 3? What about wind storms. Why are windstorms not masked? Please also consider masking regions with few events per season. There is a mask for blocking. Please state why.
  Response to R2 - comment 8:
  Thank you for the remark, there should indeed be a mask for cyclones in Fig. 3, we updated the figure. The reason why cyclone frequencies are masked, is because they are derived from the mean-sea-level pressure. Over higher terrain, this quantity has to be extrapolated from the elevated surface pressure to sea-level. This extrapolation is inaccurate over very high terrain which would lead to the identification of artificial cyclones, therefore cyclones identified over those areas are excluded from

the tracking (Murray and Simmonds, 1991). The windstorms, however, are computed from the 98th percentile of surface wind speeds, which is not influenced by high terrain. Therefore, the windstorm frequencies need no mask. For the blocking, there was no mask used. As explained in chapter 2.2 (p.6, l.7) anticyclones are only identified between 35° and 80°N. For this quantity, subtropical regions are usually excluded from blocking identification analyses to avoid the influence of the subtropical belt of high pressure systems.

- R2 - comment 9:
  The discussion is not critical enough. The reader gets the overall impression of a nearly perfect prediction system regarding the analyzed quantities. Mentioning the correlation value could sometimes already be enough to put the results in perspective. There are some inconsistencies as mentioned above that should be discussed. There is only one sentence P16, L15ff with reference to previous studies with similar objectives. Please add some references or state the lack thereof. See also point 4)
  Response to R2 - comment 9:
  We acknowledge that we have strongly emphasized the positive effects of the increased model resolution, partly at the expense of fair balance. We thoroughly double checked the discussion and rephrased expressions that could lead to the impression that HR is the perfect prediction system.
  We stated in the introduction (p.3 l.21), that our study is the first that explores the effects of model resolution on the decadal prediction skill of extra-tropical circulation metrics. However, we now added this information also to the discussion and inserted the following:
  "Although there are yet no other studies on this topic with respect to decadal time scales, our results are in agreement with findings from seasonal prediction studies (Prodhomme2016, Befort2019), who showed skill improvements for blocking, windstorm and cyclone frequencies when the same model is used and only the resolution is increased. "
  Also, the entire chapter "Discussion" was extensively revised and now relates our findings to many other studies on decadal/seasonal prediction skill or the impact of model resolution on the dynamics of the circulation.

Minor comments:

- i) The title suggests an analysis of the entire NH. Please correct. Consider adding the word "deterministic" in the title
  Response: Thank you for the remark - we changed the title to "Improvement in the decadal prediction skill of the North Atlantic extra-tropical winter circulation through increased model resolution"

- ii) "Anomaly correlation" and "skill" are used as synonyms throughout the manuscript. Please state that deterministic skill is assessed through anomaly correlation somewhere in the paper and in the abstract.
  Response: "Significant positive anomaly correlation" and "Skill" are

used as synonyms. This is stated at p.10 l.30, and a respective note was added to p.1 l.6: "The deterministic predictions are considered skillful, if the anomaly correlation is positive and significant."

- iii) There is no reference for the "common shortcoming of climate models" of a too zonal stormtrack in the introduction.
  Response: The reference Scaife et al. (2011) was added to p.3 l.5.

- iv) P1L1: The acronym MiKlip is not explained
  Response: The full name for the acronym was added to p.2 l.10.

- v) P1L8: "functional chain" is not clear
  Response: Replaced "functional chain" with "chain".

- vi) P1L11ff: Newfoundland is not "downstream" of the stormtrack.
  Response: Newfoundland is enumerated together with Central Europe, those are the regions where the windstorm frequency improves. Central Europe is downstream of the stormtrack. Newfoundland is mentioned for reasons of completeness. Though the formulation is imprecise it is not wrong. We added "primarily" to the preceding sentence, to improve precision.

- vii) P1L20ff: Please add reference for this paragraph
  Response: We added: Leckebusch2004, Ulbrich2009, Sillmann2009, Pfahl2012, DeutscheRueck2018

- viii) P2L8: "sectors" is most likely the wrong word
  Response: Replaced "sector" with " division".

- ix) P2L20ff: restructure sentence: "One result..."
  Response: Sentence was restructured.

- x) P3L1: specify "lower resolution"
  Response: about 1.5° horizontal grid spacing or less

- xi) P3L8: "functional chain?"
  Response: Replaced "functional chain" with "chain".

- xii) P3L22: change "variables" to "diagnostics" or similar. Variable is not the correct word.
  Response: Thank you for this note. We replaced "variable" with either "quantity" or "diagnostic" in various positions of the manuscript.

- xiii) P3L29: same, please check the entire manuscript
  Response: see above

- xiv) P4L12ff: "However...": Please rephrase
  Response: Rephrased to: "However, there exists no gridded observational dataset for the metrics of interest."

- xv) P5L1: add “deterministic”. See points i) and ii)
  Response: We added “deterministic”.

- xvi) P3L2: centered or uncentered anomaly correlation? See point 1)
  Response: The definition of the centered and uncentered anomaly correlation, e.g. as in Wilks’ “Statistical Methods in the Atmospheric Sciences”, refers to spatial correlations, i.e. of pairs of grid points in the observed and forecast fields. However, in our study, as described in response to R2 comment 1 and R1 comment 3, we apply a temporal correlation (Pearson) of anomalies for each individual grid point. In order to avoid misunderstandings, we could have changed the expression from “anomaly correlation” to some other, more distinct and probably longer term. But we decided to keep it like that, to be conform with previous studies of the MiKlip decadal prediction system, that also used the term ”anomaly correlation”. Also, we think the updated and very detailed description of our evaluation procedure is clear enough to avoid a misunderstanding.

- xvii) P6L32ff: This is unclear and probably wrong somehow. What kind of percentile is used? Is it the same one in LR and HR? This might explain why the difference in cyclone frequency is not apparent in windstorms
  Response: To be more clear we refined wording and replaced hindcast by model simulation. The explanation is correct. For each simulation (LR, HR, reanalysis), a different threshold is used, i.e. the local percentile of the individual simulation. This is a feature of the algorithm which implicitly adjusts means bias. The idea of the methodology is explained by Leckebusch et. al (2008). To calculate model consistent percentiles, uninitialized simulations of LR and HR are used as done by Kruschke et al (2016).

- xviii) P7L2ff: change “nicely illustrated”
  Response: Changed to “demonstrated”.

- xix) P7L31: the value in brackets is easily misunderstood. Maybe: -3% of a total of X% days in one season
  Response: Added unit information to p.7 l.31: “The blocking frequency shows a strong negative bias of fraction of blocked days per winter (-3%) in the LR system”

- xx) P10L18: change “implying” to “could be due to” or similar
  Response: Changed to “possibly due to”.

- xxi) P12L4ff: see point 2 for the whole paragraph
  Response: Revised where needed.

- xxii) P12L22ff: see point 2
  Response: We assume you mean page 13 instead of page 12. In P13L22 we simply state that areas of skill improvement coincide with areas of bias improvement. There is no description of dependency or of cause and effect.

- xxiii) P13L35ff: improvement in cyclone frequency improves windstorm frequency? Specifically along the European western coast? P10L12ff highlights the differences of the 2 diagnostics
  Response: The differences between windstorms and cyclones explained in P10L12ff refer to the positive cyclone frequency bias over the central North Atlantic, which is caused (as we had suggested in the submitted draft and now proved in the review process) by weak and short-lived cyclones. The argumentation in this paragraph is used to clarify that a rather weak bias in the windstorm frequency is not at all contradictory to the strong cyclone frequency bias, because the windstorms can be considered a subset of the cyclones, and the other subset which is not equivalent to the windstorms (i.e. the weak cyclones) can explain the positive cyclone frequency bias. This explanation is not contradictory to the fact that the skill in cyclone frequency affects the skill of the windstorm frequency (P13L35ff), because still the cyclone frequency covers all intensities of systems, those that do and those that do not produce storms. It is therefore possible and likely, that the subset of strong cyclones influences the windstorm frequency and its skill, respectively.

- xxiv) P15L8: Muller et al 2018 show a decrease of MSLP bias in the Eastern North Atlantic but an increase in the Western North Atlantic in HR. It is therefore only partially "in line".
  Response: We added "over the central North Atlantic" to be more precise.

- xxv) P15L25ff: see point 2, for blocking + cyclones
  Response: Again, we only say the areas of skill and bias improvement coincide. We rephrased the second part of the sentence.

- xxvi) P5L10: Please provide a reference or calculation method for the statistical significance. Is the calculation method different between correlation significance and significance for the differences in correlation
  Response: A reference was added (Goddard et al., 2013) to p.5 l.4.

2 List of all relevant changes made in the manuscript

- two words of the title were changed in agreement with the editor
- paragraph for the description of the computation of the anomaly correlation was added
- orography mask for cyclone frequency in Fig. 1, 5 was corrected
- orography mask for cyclone frequency in Fig. 3 was added
- figure for the distribution of the strength and lifetime of North Atlantic cyclones was added
- discussion and conclusions were separated, discussion was supplemented by relating our findings to other studies

3 Marked-up manuscript version

Improvement in the decadal prediction skill of the ~~northern hemisphere~~ North Atlantic extra-tropical winter circulation through increased model resolution

Mareike Schuster[1], **Jens Grieger**[1], **Andy Richling**[1], **Thomas Schartner**[2], **Sebastian Illing**[1], **Christopher Kadow**[1], **Wolfgang A. Müller**[4], **Holger Pohlmann**[3,4], **Stephan Pfahl**[1], **and Uwe Ulbrich**[1]

[1]Freie Universität Berlin, Institut für Meteorologie, Carl-Heinrich-Becker Weg 6-10, 12165 Berlin
[2]Deutscher Wetterdienst, Güterfelder Damm 87-91, 14532 Stahnsdorf
[3]Deutscher Wetterdienst, Bernhard-Nocht-Straße 76, 20359 Hamburg
[4]Max-Planck-Institut für Meteorologie, Bundesstraße 53, 20146 Hamburg

Correspondence: Mareike Schuster (mareike.schuster@met.fu-berlin.de)

Abstract. In this study the latest version of the MiKlip decadal hindcast system is analyzed and the effect of ~~different~~ an increased horizontal and vertical ~~resolutions~~ resolution on the prediction skill of the ~~northern hemisphere~~ extra-tropical ~~atmospheric~~ winter circulation is assessed. Four different metrics - the stormtrack ~~, blockingfrequencies, cyclone frequencies~~ as well as blocking, cyclone and windstorm frequencies - are analyzed ~~with respect to the anomaly correlation of their winter averages~~in the North Atlantic and European region. The model bias and the deterministic decadal hindcast skill are evaluated ~~in both,~~ for an ensemble of 5 members in each a lower resolution version (LR, atm: T63L47, ocean: 1.5° L40) and a higher resolution version (HR, atm: T127L95, ocean: 0.4° L40) of the MPI-ESM system~~,~~. The skill is assessed for the lead ~~years~~ winters 2-5 in terms of the anomaly correlation of the quantities' winter averages, using initializations between 1978 and 2012. The deterministic predictions are considered skillful, if the anomaly correlation is positive and statistically significant. While the LR version shows common shortcomings of lower resolution climate models, e.g. a too zonal and southward displaced stormtrack and a negative bias of blocking frequencies over the eastern North Atlantic and Europe, the HR version ~~works against~~ counteracts these biases. As a result, a ~~functional~~ chain of significantly improved decadal prediction skill between all four metrics is found with the increase of the spatial resolution. While the ~~stormtrack,~~ skill of the stormtrack is significantly improved primarily over the main source region of synoptic activity - the North Atlantic Current ~~, -~~ the other extra-tropical ~~measures~~ quantities experience a significant improvement ~~downstream thereof~~primarily downstream thereof, i.e. in regions where the synoptic systems typically intensify. Thus, the skill of the cyclone frequencies is significantly improved over the central North Atlantic and Northern Europe, the skill of the blocking frequencies is significantly improved over the Mediterranean, Scandinavia and Eastern Europe and the skill of the windstorms is significantly improved over Newfoundland and Central Europe. Not only is the skill improved with the increase in resolution, but the HR system itself exhibits significant skill over large areas of the North Atlantic and European sector for all four circulation metrics. These results are particularly promising regarding the high socio-economic impact of European winter windstorms and blocking situations.

1 Introduction

The extra-tropical circulation plays an important role for ~~extremes in~~ the redistribution of energy in the atmosphere. The prevailing westerlies and the embedded cyclones and anticyclones determine the weather and climate ~~, as its variability determines~~ of the mid-latitudes, assisting in balancing temperature and humidity contrasts between tropical and polar regions. Natural climate variability as well as externally forced climate change determine fluctuations in the circulation and thus i.a. the frequency of ~~extreme cyclones, embedded storm fields and~~ extremes such as strong cyclones, intense windstorms or phases of blocked flow. The consequences of such features include extremes in temperature, precipitation/drought and wind speed, often accompanied by immense damage and harm (e.g. ?????). Therefore, the societal demand for reliable near term climate predictions of such features - also to support political, economical and administrative decision making - is perpetually growing.

Decadal Climate prediction itself is an active research field in climate sciences. Different research groups around the globe aim at the development of skillful prediction systems (Boer et al., 2016). Retrospective forecasts, termed hindcasts, are used to assess the ability of the model systems to predict climate variability on inter-annual to decadal time scales. Initialized from observation-based data and run for a period of 10-30 years, decadal climate predictions combine forecast elements from weather and seasonal forecast ~~sectors~~ divisions (initial conditions) as well as from long-term climate projections (boundary conditions). To date, different designs of decadal prediction systems are prevalent. They either consist of a multi-member single-model suite, such as the UK MetOffice's Decadal Prediction System 'DePreSys' (Smith et al., 2007) and the German '~~MiKlip~~ Mittelfristige Klimaprognosen (MiKlip) system' (Marotzke et al., 2016)~~–~~, or they are based upon a multi-model suite, as used e.g. within the 5th Coupled Model Intercomparison Project (CMIP5; Taylor et al. 2012). Several multi-model studies, using the CMIP5 decadal prediction suite, come to the conclusion that there exists significant prediction skill on decadal time scales (e.g. Kim et al., 2012; Doblas-Reyes et al., 2013; Meehl et al., 2014). Results from these studies have also been included in the 5th assessment report (AR) of the Intergovernmental Panel on Climate Change (IPCC; Kirtman et al. 2013). Currently, in preparation for the 6th AR of the IPCC and CMIP6, improved decadal prediction systems are developed (Eyring et al., 2016; Boer et al., 2016; Kushnir et al., 2019).

With respect to atmospheric quantities, various versions of the 'MiKlip system', based on the Max-Planck Institute Earth System model (MPI-ESM), show decadal prediction skill mainly in terms of global and regional temperature indices ~~Müller et al. (2012); Pohlmann et al. (2013); Kröger et al. (2017)~~(Müller The skill for precipitation, due to its complex and partly small-scale nature, is however regionally confined and prevalently limited to lead year 1 (Kadow et al., 2016). ~~One result of MiKlip ('Mittelfristige Klimaprognosen', German for medium-term climate predictions)~~ Results of practical relevance ~~is the skillful~~ within MiKlip are on the one hand skill for the prediction of 10 m wind speed and wind energy over Central Europe, ~~that~~ which exists irrespective of the ocean initialization technique (Moemken et al., 2016) and ~~for~~ on the other hand skill for the prediction of wind speeds of different quantiles upward from 75% (Haas et al., 2016).

Kruschke et al. (2014) and Kruschke et al. (2016) analyzed the forecast skill of Northern Hemisphere cyclone and windstorm frequencies in the ~~'MiKlip system1000~~MiKlip system, respectively. They found probabilistic decadal forecast skill for cyclone frequencies in some areas over the Northern Hemisphere ocean basins, mainly the European North Sea and the central Pacific, ~~with~~ whereupon the sub-sample of strong cyclones ~~exhibiting~~ exhibits generally higher skill than the complete sample of all cyclones (Kruschke et al., 2014). Using a parametric bias adjustment approach, Kruschke et al. (2016) found windstorm frequencies for winters 2–5 and winters 2–9 to be skillful over large parts of the Northern Hemisphere when compared against climatological forecasts. Another study using the MPI-ESM decadal prediction system demonstrated that the decadal prediction skill for surface temperature and cyclone frequencies can be significantly improved by replacing each ensemble member's ocean state with the ensemble mean ocean state at regular intervals during the forecast period (Kadow et al., 2017). More studies using the MPI-ESM hindcasts are collected in a special issue of the Meteorologische Zeitschrift about the validation of the MiKlip system in its first phase (Kaspar et al., 2016).

~~With~~ It is well known that a coarse spatial resolution of global coupled climate models hinders the proper representation of sub-synoptic scale systems, and thus the climate mean state and variability. For example, with respect to the North Atlantic and European domain ~~, in many climate models of lower resolution~~ many lower resolution climate models exhibit a cold sea surface temperature (SST) bias south of Greenland~~is present~~, due to a displacement of the North Atlantic Current or a too weak overturning circulation (Park et al., 2016; Scaife et al., 2011; Wang et al., 2014). This common bias in the North Atlantic Current ~~,~~is associated with a too zonal stormtrack, stronger geopotential height gradients in the mid-latitudes, increased westerlies and reduced blocking frequencies over Europe (e.g. Scaife et al., 2011). It has been found in many studies ~~,~~that the atmospheric dynamics benefit not only from a coupling of the atmosphere and ocean but also from an increased model resolution ~~(e.g. Shaffrey et al., 2009; Jung et al., 2012; Dawson et al., 2013)~~(e.g. Sh Scaife et al. (2011), for example, demonstrated that the increase in resolution in both the atmospheric and oceanic model components results in a ~~functional~~ chain of improve-

ments, as they found a reduced SST bias in the higher resolution model which in turn lead to a better representation of westerly winds and blocking frequencies. Similar effects can also be found in atmosphere only models, e.g. for the blocking frequency bias in **?**.

With an atmospheric resolution of T63L47 (about 1.875° horizontal grid spacing) and an oceanic resolution of 1.5°L40 the MPI-ESM-LR decadal prediction system ~~,~~ applied in the first phase of MiKlip ~~,~~ has a rather moderate spatial resolution. Meanwhile, studies using higher resolution forecast systems are available, for instance Monerie et al. (2017) using 0.5° grid spacing in the atmosphere and 0.25° in the ocean and Robson et al. (2018) using ∼0.9° in the atmosphere and 0.25° in the ocean. They focus on oceanic parameters and find skill e.g. for SSTs, sea ice extent and ocean heat content, respectively. However, systematic analyses of the actual effect of the increase in resolution on the hindcast performance on the decadal scale are rare. Pohlmann et al. (2013) found for the hindcasts of mixed resolution (MPI-ESM-MR) that an increase in vertical (atmosphere: T63L95) and horizontal resolution (ocean: 0.4°L40) compared to MPI-ESM-LR improves the tropical Pacific surface temperature predictions in the lead years 2-5 and leads to a good representation of the quasi-biennial oscillation (QBO), which remains in alignment with observations well beyond the first 12 months after initialization. Apart from that, the mixed resolution shows only modest benefit for the hindcast skill (Marotzke et al., 2016).

In this study, for the first time an analysis of the direct impact of the model resolution on the skill of decadal climate predictions of dynamical variables is performed under otherwise unchanged model settings (parametrization and initialization)~~is performed~~. We evaluate the MiKlip hindcasts performed with the latest version of the Max-Planck Institute Earth System model with higher resolution (MPI-ESM-HR, Müller et al. 2018), which will contribute to CMIP6, and compare its decadal forecast skill to that of a previous lower-resolution version (MPI-ESM-LR). While many studies analyzing the skill of decadal forecast systems tend to focus on basic atmospheric variables such as the surface temperature and precipitation (e.g. Smith et al., 2007; Keenlyside et al., 2008; Goddard et al., 2013; Kadow et al., 2016; Monerie et al., 2018; Xin et al., 2018), we emphasize the role of dynamical processes and therefore analyze a set of ~~variables~~ quantities representing the extra-tropical winter dynamics~~–~~: the stormtrack, blocking, cyclones and windstorms.

We introduce the MPI-ESM prediction system ~~in Sect. 2.1,~~ as well as the skill measure used to assess the hindcast quality in ~~this study~~Sect. 2.1. In Sect. 2.2 we describe the different circulation quantities in detail and present their climatology in the ERA-Interim reanalysis with a focus on the North Atlantic and European region. The model ~~climatology~~ climatologies and biases are discussed in Sect. 3.1 ~~. Finally,~~ and the prediction skill of the winter circulation is evaluated in Sect. 3.2~~with a focus on the North Atlantic and European region~~. In Sect. 4 we discuss and relate our findings to other studies, before we conclude our results in Sect. 5.

2 Data and methodology

The extra-tropical circulation in the Northern Hemisphere is most active during the winter season, with a stronger jet stream in the upper-troposphere and numerous strong cyclones developing in the mid-latitude baroclinic areas, favored by strong horizontal temperature contrasts resulting from relatively warm ocean currents near the surface and cold polar air masses. Storms that strike the European continent at this time of the year are often powerful and damaging. We will therefore focus on the winter circulation and evaluate averages of the stormtrack and blocking, cyclone and windstorm frequencies from October through March. The stormtrack describes the variability of baroclinic waves on synoptic time-scales in the extra-tropics. These baroclinic waves are a combination of two contributing components, i.e. anticyclonic and cyclonic anomalies, which we will analyze in terms of blocking frequencies on the one hand and extra-tropical cyclone and windstorm frequencies on the other hand.

To assess the model bias and to compute the prediction skill of the different ~~variables~~ diagnostics in the decadal hindcasts, a reference ~~,~~ (i.e. ~~an observational~~observational) data set is needed. However, there exists no gridded observational ~~data set of the dynamical variables that we aim at~~dataset for the metrics of interest. Instead we make use of a reanalysis product and derive the circulation quantities for the winters 1979/80 to 2016/17 from the ERA-Interim reanalysis (Dee et al., 2011), created by the European Centre for Medium-Range Weather Forecasts (ECMWF), with a horizontal resolution of T255 (∼0.75°) on 60 levels and a top of the atmosphere at 0.1 hPa.

2.1 Forecast system and skill measures

The two decadal forecast systems that we compare are both based on the Earth System Model of the Max-Planck-Institute for Meteorology (MPI-ESM) version 1.2, which is a coupled atmosphere ocean model and consists of the atmospheric component ECHAM6.3 and the oceanic component MPI-OM1.6.2.

The lower resolution of MiKlip's pre-operational decadal prediction system (MPI-ESM-LR, termed LR hereafter) has an atmospheric horizontal resolution of T63 (1.875°) and 47 levels, with the top of the atmosphere at 0.01 hPa (Mauritsen et al., 2019). The ocean component is run with 1.5° L40. A general skill assessment of decadal predictions performed with the LR system can be found in Polkova et al. (2019). The higher resolution version (MPI-ESM-HR, termed HR hereafter) uses T127 (0.9375°) and 95 vertical levels for the

atmosphere, and 0.4° L40 for the ocean (Müller et al., 2018). The HR version therefore has a finer grid in both the atmosphere and the ocean components. For this analysis, both systems use the CMIP5 external forcing with respect to greenhouse gases and aerosols (for details see Giorgetta et al. 2013). Both systems are full-field initialized in the atmosphere, using ERA-40 (Uppala et al., 2005) and ERA-Interim (Dee et al., 2011); and anomaly-initialized in the ocean, using ORA-S4 (Balmaseda et al., 2013) and sea-ice concentration from the National Snow and Ice Data Center (NSIDC). The initialization procedure is identical to the one used for MiKlip's Baseline 1 system and is described in more detail in Pohlmann et al. (2013). The LR system consists in total of 10 ensemble members, initialized annually between 1960 and 2016, with each initialization covering one decade. The integration period for each of the initializations spans 10 years. However, since the HR system - with an otherwise identical hindcast setup - consists of only 5 members, and to guarantee a fair comparison between the two forecast systems we only evaluate the first 5 members of LR as well.

To ~~determine the~~ derive the deterministic skill of the two forecast systems, we ~~analyze the~~ focus on the temporal variability and analyze the anomaly correlation for the winters 2-5 ~~after initialization~~ (Oct-Mar), following the Decadal Climate Prediction Project (DCPP, Boer et al. 2016) protocol. ~~We focus on the temporal variability by analyzing the~~ That means that we calculate lead time dependent anomalies of the circulation measures. This is a simple and robust approach to account for a possible lead time dependent mean bias, i.e. drift. Thus, for each of the initialization experiments (1978, 1979, ...) the ensemble average (5 members) of the temporal mean of the 4 contained lead winters is calculated per grid point. This forms a new ensemble mean time series of the lead winters 2-5. This time series serves to calculate the climatology (temporal mean) as well as the respective anomaly time series. The time series of those anomalies of the hindcasts is then correlated (Pearson) to the time series of anomalies of the reanalysis. In decadal prediction studies, this procedure is usually repeated for each lead time, e.g. lead year 1, lead year 2-5, lead year 6-9 - it is therefore referred to as lead time dependent anomaly correlation. ~~Therefore, we employ~~ In our study we only show results for one lead time: lead winters 2-5. The initialization of the hindcasts takes place in October, this means the first full winter that we analyze is the second winter, i.e. the months 12-17 (Oct-Mar) after initialization. This evaluation procedure is part of the decadal climate prediction evaluation software ~~of~~ that was designed within the MiKlip project (Illing et al., 2014) ~~.~~ and is applied for this study. This OpenSource evaluation software follows the evaluation framework of Goddard et al. (2013) which led to the DCPP requirements.

To match the period covered by the ERA-Interim reanalysis, we do not use the full set of initializations but instead use the decadal hindcast experiments that are initialized between 1978 (winter 2: 1979/80, winter 5: 1982/83) and 2012 (winter 2: 2013/14, winter 5: 2016/17) in LR and HR. In total we therefore analyze 700 October-to-March winter seasons (5 members x 35 initializations x 4 lead ~~years~~winters) per forecast system. The skill of each of the forecast systems (LR, HR) is first evaluated against the reanalysis data, i.e. the anomaly correlation between the respective hindcast and ERA-Interim is determined. Then, the two systems are compared against each other, i.e. the difference of the aforementioned correlations between the two forecast systems is computed. To determine the significance of the correlation (95% significance level), the time series of reanalysis-hindcast pairs is resampled with replacement 1000 times (block bootstrap taking auto-correlation into account), following Goddard et al. (2013).

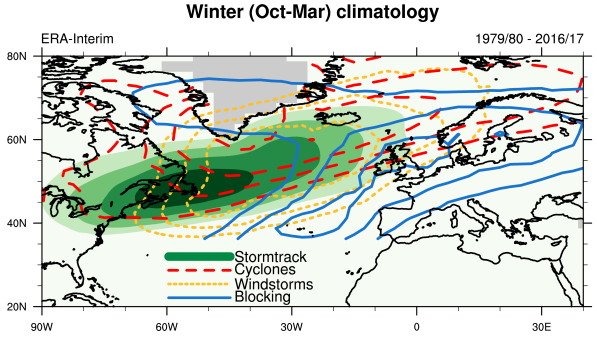


Figure 1. Climatology of the winter average (Oct-Mar) of different circulation quantities in the ERA-Interim reanalysis for the period 1979/80-2016/17. The stormtrack, i.e. the standard deviation of the 500 hPa geopotential height anomaly is shown in m (45-60 by 5). The fraction of blocked days is shown in % (4-8 by 2). The cyclone frequency (120-180 by 20) and windstorm frequency (25-30 by 2.5) are shown in number of tracks within a radius of 1000 km. Grey masked areas denote grid points with an orography larger than 1500 m, which have been omitted for cyclone identification.

2.2 Circulation metrics

Stormtrack

The extra-tropical stormtrack is derived from the bandpass filtered variability of the geopotential height ~~field~~ at 500 hPa in the window of 2.5 to 6 days ~~-~~, an Eulerian approach following Blackmon et al. (1976). Its long term winter average (October through March) is displayed in Fig. 1 for the North Atlantic and European region and the period 1979/80-2016/17 based on the ERA-Interim reanalysis. The North Atlantic stormtrack is visible in green shades, with its maximum of ~~60m~~ 60 m located over the western North Atlantic and Newfoundland and a typical north-eastward tilt.

Blocking

~~The second synoptic scale feature that we analyze is atmospheric blocking . Here,~~ For atmospheric blocking a

slightly modified version of the 2-dimensional blocking index of Scherrer et al. (2006), based on gradients in the daily 500 hPa geopotential height field, is used to identify instantaneously blocked grid points. In contrast to Scherrer et al. (2006), where a blocking area is defined in between the blocking high and the associated low, here the position of detected blocked grid points is shifted north by 7.5° to correspond better with the anticyclonic part of a blocking situation. To account for large-scale and persistent blocking anticyclones between 35°N and 80°N, an adapted tracking algorithm for blocking regimes, similar to the approach by Barnes et al. (2012), is applied. With this tracking method, we only select contiguously blocked regions with a minimum zonal and meridional extension of ∼15° and an area of at least 1.5 x 10^6 km^2 lasting for a minimum of 4 days. A possible shifting, merging and splitting of blocking areas in time is considered by adopting a blocking overlap area criterion of 750.000 km^2 between two consecutive days and a maximum distance between blocking centers of 1000 km. The climatology of the mean winter blocking frequency is displayed in blue isolines in Fig. 1. Its maximum of 8% blocked days stretches from the Azores to Scotland. A second region of increased blocking frequencies is found between Greenland and Iceland.

Cyclones

To identify and track extra-tropical cyclones we apply an objective Lagrangian feature tracking algorithm, developed by Murray and Simmonds (1991)~~to~~ , to 6-hourly values of the mean sea level pressure. Maxima of the Laplacian of the mean sea level pressure are identified and, if a minimum in the pressure field itself can(not) be detected in the vicinity, a closed (open) cyclone is identified. The system is then tracked in time, at 6-hourly time steps~~, using predicted locations for the successive time step, and probabilities for the assignment of the systems in the consecutive time steps~~. Only cyclones that live for more than 24 hours and reach a Laplace larger than 0.7hPa/(degree latitude)^2 and have closed isobars at least once during their lifetime are selected for evaluation. The measure we ultimately use for our evaluation is the cyclone frequency, i.e. the number of cyclone tracks that pass within a radius of 1000 km of the respective grid point on a 2.5° x 2.5° grid. As the extrapolation of pressure to sea level can be erroneous over high terrain, cyclones are not identified at grid points where the orography is higher than 1500 m. ~~In this study only cyclones are taken into account that are strong (Laplacian $>$ 0.7 (deg. lat.)$^{-2}$) and closed at least once during their lifetime and that last longer than a day.~~ The winter average of the cyclone frequency is displayed in Fig. 1 ~~as~~ in red dashed contours. Its maximum is located at the southern tip of Greenland with 180 cyclones and a band of enhanced cyclone frequencies is located downstream of the stormtrack maximum with a similar southwest-northeast tilt.

Windstorms

Yet another objective Lagrangian tracking scheme is used to derive the frequency of extra-tropical windstorms (Leckebusch et al., 2008; Kruschke, 2014). This method is based on the exceedance of the local 98th percentile of the near-surface wind speed to define contiguous fields of strong wind. Percentiles are calculated for each ~~hindcast~~ model simulation (LR, HR) and the reanalysis individually~~using 6 hourly~~ , using 6-hourly data of the whole year between 1981 and 2010. For the hindcasts the percentiles of the uninitialized counterparts are used as done by Kruschke et al. (2016). Windstorms are identified if the area of wind exceedance above the percentile is larger than 150.000 km^2 and if the feature is trackable for at least 18 hours. Tracking is done by means of a nearest neighbour approach. ~~Windstorm~~ The individual windstorm tracks are further used to calculate windstorm frequencies, which are computed identically to those of the cyclone frequencies. The yellow dotted contours in Fig. 1 represent the average winter windstorm frequency. ~~It is nicely illustrated~~ Its maximum of 30 windstorms is located also downstream of the stormtrack maximum, but slightly shifted southward compared to the cyclone frequencies. This illustrates that the corresponding windstorm field is usually located to the south of the cyclone center, where the pressure gradients and thus geostrophic wind velocities are typically largest.~~Its maximum of 30 windstorms is located also downstream of the stormtrack maximum, but slightly shifted southward compared to the cyclone frequencies.~~

The software ~~that was~~ routines that were used to compute all the extra-tropical circulation quantities, as well as the evaluation procedure were implemented as separate plug-ins into the MiKlip Central Evaluation System (https://www-miklip.dkrz.de) - based on the Free Evaluation System Framework (Freva, Kadow et al. in preparation) - by their developers and authors of this paper. The single plug-ins and their documentation can be found under https://www-miklip.dkrz.de/plugins - plus the respective suffix /stormtrack/detail/ (Stormtrack); /blocking_2d/detail/ (Blocking); /zykpak/detail/ (Cyclones); /wtrack/detail/ (Windstorms) and /murcss/detail/ (skill analysis).

3 Results

3.1 Model bias

Before we analyze the decadal prediction skill, we will first evaluate the ensemble mean climatology and ~~the model~~ bias, in order to assess the model's capability to represent the four atmospheric circulation features. For this, we only take into account those seasons that will be used for the skill analysis, i.e. the winters 2-5 of each of the 35 initializations (1978-2012) and ~~2x~~ 2 x 5 members (LR~~and~~ , HR). To compute the model bias, we consider the entire reanalysis

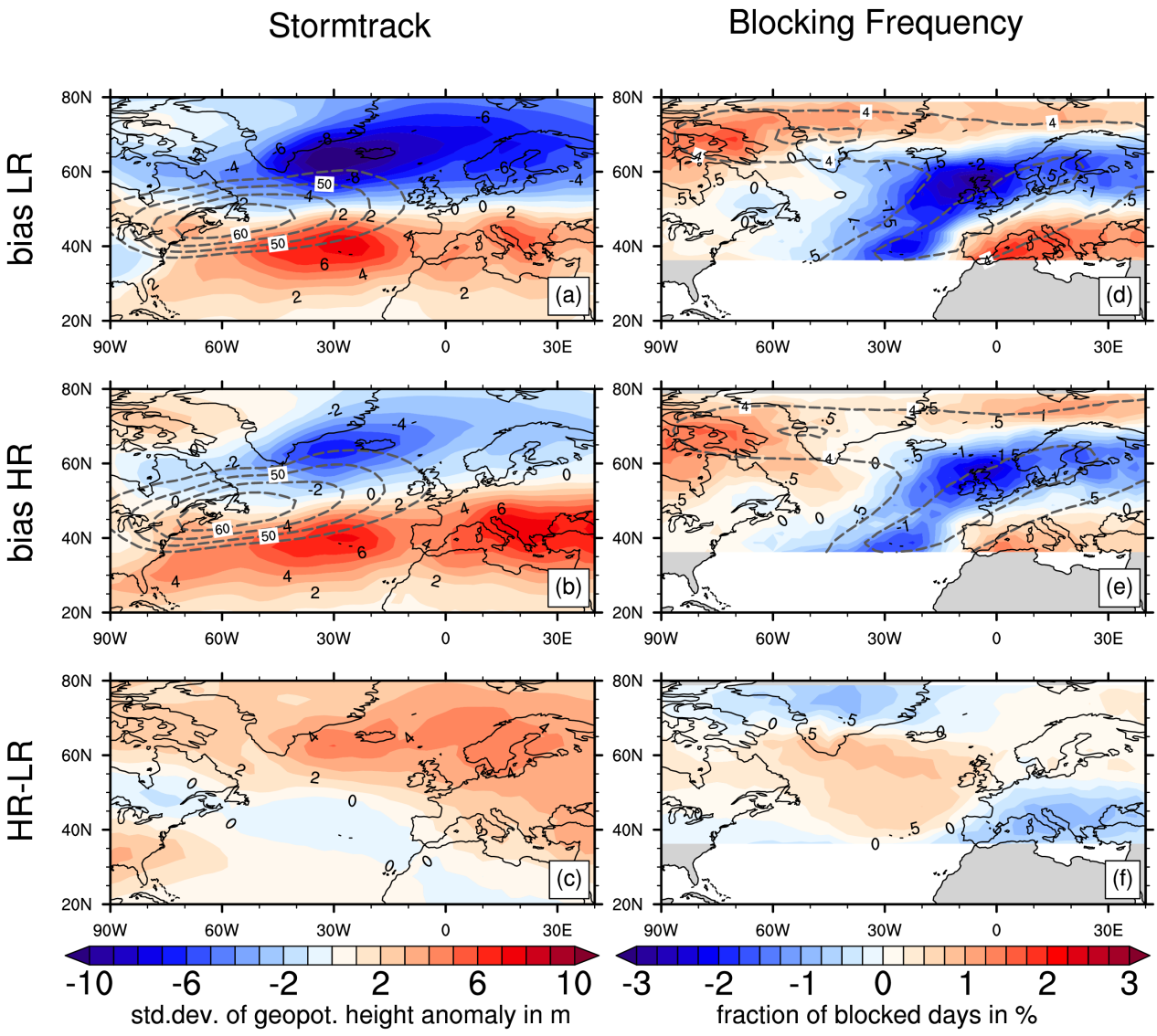


Figure 2. Ensemble mean model bias relative to ERA-Interim (shading) and model climatology (dashed contours) of the respective circulation quantity in LR (top row), HR (middle row) and the difference between HR and LR (bottom row). The circulation quantities displayed are the stormtrack (left) and the blocking frequency (right). Initializations from the period 1978-2012 are used for 5 members of each, LR and HR, and the ensemble mean is computed from lead-time averages over the hindcast winters 2-5 (Oct-Mar). In ERA-Interim the winters between 1979/80 and 2016/17 are used. The grey contours, i.e. ensemble mean climatology, have the same levels as in Fig. 1 - 45-60 by 5 m for the stormtrack and 4-8 by 2 % for the blocking frequency.

data set, i.e. winters from 1979/80 to 2016/17.

In Fig. 2 the model bias ~~, compared to ERA-Interim,~~ for the stormtrack and blocking frequency compared to ERA-Interim is displayed in colored shades and the respective model climatology is shown in grey contours, for both the LR and HR ensemble mean. The grey contour levels are the same as for the ERA-Interim climatology in Fig. 1. The LR system shows the typical North Atlantic stormtrack along 45°N, with a maximum over the western part of the basin, however rather zonally aligned and shifted southward (Fig. 2a). Since the observed stormtrack is tilted ~~,~~ from south-west to north-east (see Fig. 1), this results in a negative bias (-10m) at higher latitudes and a positive bias (+8m) at lower latitudes in the LR prediction system. This bias can partly be corrected with the increase in the model resolution, as the HR system increases the stormtrack activity where there is a negative bias in LR and vice versa (Fig. 2c), however this effect is strongest at the northern side of the stormtrack, as also seen in Müller et al. (2018)~~(their Fig. 10)~~. In HR the North Atlantic stormtrack is more tilted, and therefore closer to observations (Fig. 2b). Not only does

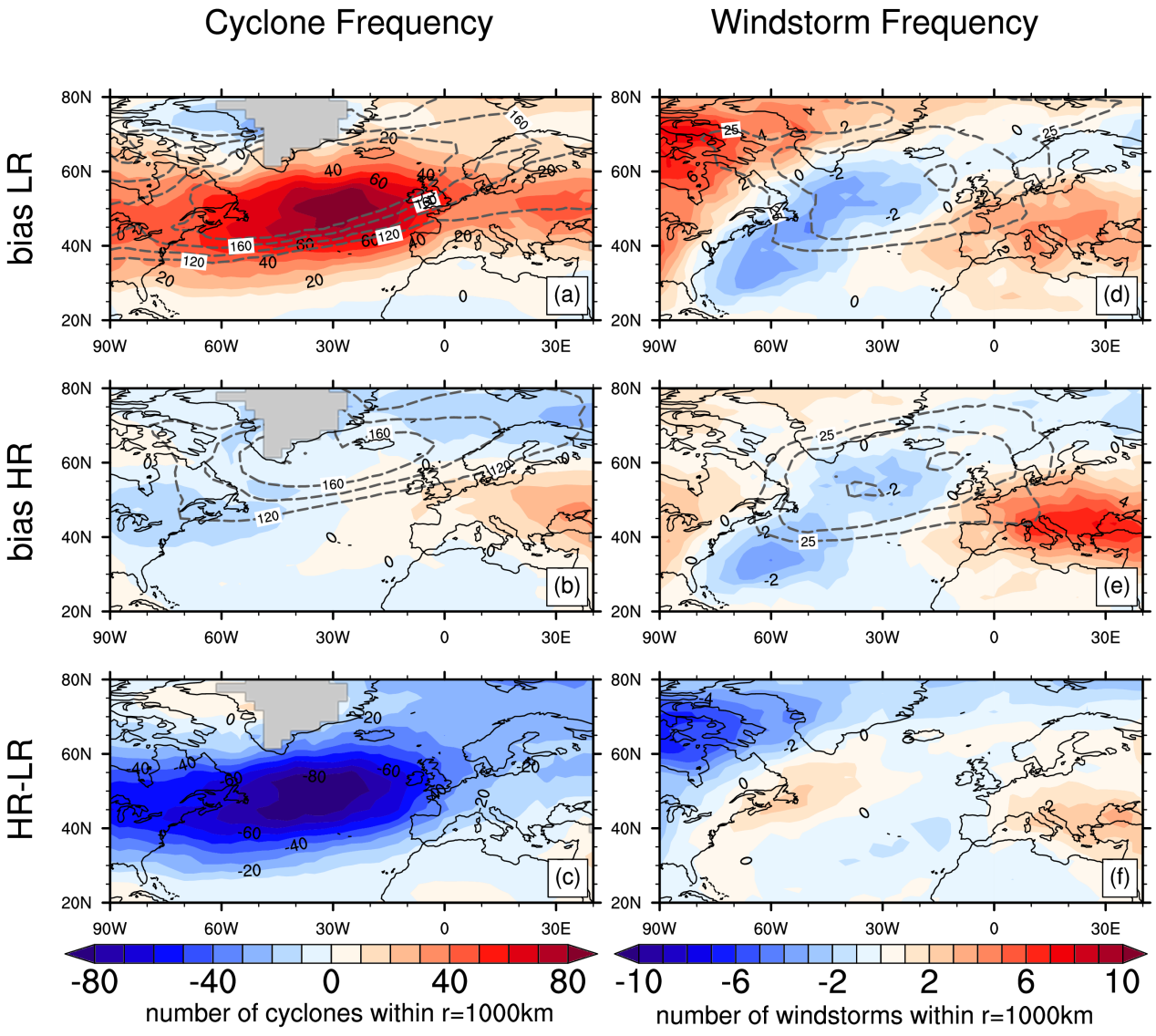


Figure 3. Same as Fig. 2 but for the cyclone frequency (left) and the windstorm frequency (right). The grey contours, i.e. ensemble mean climatology, have the same levels as in Fig. 1 - 120-180 by 20 cyclones for the cyclone frequency and 25-30 by 2.5 storms for the windstorm frequency.

it extend further north in the higher resolution system, but it also extends further downstream ~~,~~ towards Central and Eastern Europe, and therefore reduces the negative bias over the North Sea and Scandinavia that is present in LR. The bias in HR is reduced at both ~~,~~ the northern and southern ~~,~~ flanks of the Atlantic stormtrack, however ~~it~~ the southward shift over the central North Atlantic is still present (-7m and +7m).

The blocking frequency shows a ~~strong negative bias~~ negative bias of fraction of blocked days per winter (-3%) in the LR system ~~,~~ just north of its climatological maximum, i.e. over a band stretching from the central North Atlantic and Great Britain towards the Baltic Sea, and a positive bias (+1.5%) over the Mediterranean (Fig. 2d). Fig. 2f ~~nicely~~ illustrates that again the HR prediction system counters these shortcomings of the LR system and reduces the bias in the right places, but the effect is rather marginal for this quantity. Though weaker, the bias of the blocking frequency in HR is still considerable (-2.5% and +1% respectively). These findings are in line with the analysis of blocking in Müller et al. (2018)~~(their Fig. 12)~~.

The climatology of the cyclone frequency with its maximum at the southern tip of Greenland, seen in Fig. 1, is also visible in LR (Fig. 3a). In contrast to the stormtrack and blocking frequency, the cyclone frequency in LR does not exhibit a clear southward shift compared to the reanalysis. Instead, in the low resolution system there are overall far too

many cyclones present between 30°-70°N, but especially over the central North Atlantic ~~,~~ where a positive bias of up to +80 cyclones is found. Most impressively amongst all variables, this bias of the cyclone frequency is radically reduced and almost completely absent in the HR system (Fig. 3b). The numbers are reduced to a bias of -10 cyclones over the western North Atlantic and +10 cyclones over Europe. The increase in horizontal and vertical resolution evidently eliminates many cyclone tracks in the MPI-ESM (Fig. 3c) over the entire North Atlantic domain and adjacent continents. This effect is further discussed in Sect.4. This results in cyclone climatologies very close to those in ERA-Interim in HR (Fig. 3b).

The windstorm frequency shows a slightly different behavior. There are too few windstorms (-3) present over the western and central North Atlantic along the North Atlantic current, and too many windstorms over the continents - +3 over Europe and +5 over ~~the US~~ North America and Canada (Fig. 3d). Given that there are too many cyclones in LR, the negative windstorm bias over the Atlantic might seem contradictory, as windstorms are a consequence of strong cyclones. However, it should be highlighted that the cyclone tracking algorithm also detects ~~cyclones in their weakphase, as long as they become strongat least once during their lifetime. Therefore,~~ weak and moderate cyclones. Thus, the windstorms displayed in Fig. 3d-f can be considered a subset of all (weak, moderate and strong) cyclones displayed in Fig. 3a-c. This suggests that the positive cyclone ~~bias is likely influenced by many weaker and~~ frequency bias in LR is caused by weak cyclones. This is confirmed by the intensity and lifetime histograms displayed in Fig. 4, which only include cyclones that pass the central North Atlantic (50°-10° W, 40°-60°N) at some point during their lifetime. While the distributions of those North Atlantic cyclones match very well between HR and ERA-Interim, the LR prediction system overestimates weak to moderate (0.7-2.2 hPa~~/or short lived systems that are not strong for a long enough time~~ (degree latitude)^2) and short- to average-lived (1-7 days) cyclones. Those cyclones are however usually not strong enough to develop a windstorm. A similar feature was reported by Kruschke et al. (2014) for the uninitialized LR runs of the previous MPI-ESM system. Although the positive cyclone frequency bias is generally weaker for the uninitialized runs, they demonstrated that it can mainly be attributed to weak and moderate systems, by illustrating a reduced bias over the North Atlantic and Europe when only intense cyclones, i.e. the strongest 25% in terms of the Laplacian of the sea-level pressure, are considered. The negative windstorm ~~bias~~ frequency bias over the central North Atlantic is therefore not contradictory. In fact it is in line with the too zonally oriented stormtrack (Fig. 2a,b), also resulting in too many windstorms over Central Europe and the Mediterranean and too few storms over Northern Europe (Fig. 3d,e). The increase of the model resolution yields an increase of windstorm frequency over the North Atlantic current (Fig. 3e) and a remarkable reduction over the Hudson Bay~~, implying that local temperature gradients along land-sea borders and related surface fluxes, are slightly better represented in HR~~. The bias over South East Europe, however, is amplified. This leaves the higher resolution system with biases of -2 along the North Atlantic current and the central North Atlantic, and +6 over South East Europe (Fig. 3f).

While the exact location and magnitude of the extra-tropical circulation features over the North Atlantic and European region exhibits deviations from the observation, overall the MPI-ESM is capable to represent those dynamical ~~variables~~quantities. Also in Müller et al. (2018), it is noted that although bias reductions from LR to HR are modest for the multitude of ~~variables~~ diagnostics they analyzed, the dynamics of the atmosphere still benefit from the increase in resolution and make this model eligible for prediction studies. We therefore proceed to analyze the deterministic decadal prediction skill.

3.2 Prediction skill

The anomaly correlation between the stormtrack in the LR hindcast and ERA-Interim for the winters 2-5 after initialization is shown in Fig. ~~4~~5a. Although both ~~,~~ significant positive and negative correlations are equally valuable from a mathematical point of view, a significant negative correlation, i.e. a consistently opposite prediction of the observed climate variability is inconsistent with the physically-based model setup. We thus consider only significantly positive correlations as model prediction skill. The LR system shows skill for the stormtrack over the central North Atlantic (correlation coefficient r=0.4), as well as over Canada, the Baffin Bay and the Barents Sea. However, southwestward of the climatological stormtrack maximum, over the North Atlantic Current, where the meridional gradient of the stormtrack climatology is strongest, there is significant negative correlation (r=-0.3). This lack of skill in that area is overcome when the resolution of the dynamical model is increased. In the HR system (Fig. ~~4~~5c) there is a large area of significant positive correlation over the North Atlantic Current (r=0.6) and additionally over Iceland and Central Europe (r=0.5). The improvement from LR to HR, shown in Fig. ~~4~~5e, is strongest over the North Atlantic Current and the tropical North Atlantic. Additionally, there are areas of statistically significant skill improvement east of the Azores, west of Iceland and over the North and Baltic Seas. ~~The improvement over the latter two~~ Although positive anomaly correlations are not a direct consequence of bias reductions, the better representation of the average circulation and its variability does have an impact on the anomaly correlation and thus the prediction skill. Therefore, the skill improvement over those regions is in line with the

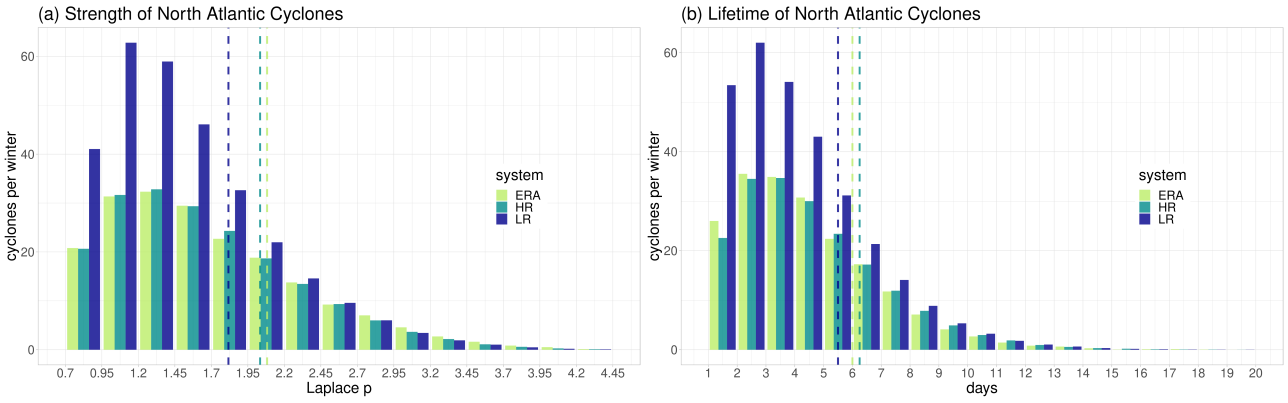


Figure 4. Histograms of a) the strength (max. along track Laplacian of the sea level pressure in hPa/(degree latitude)^2) and b) the lifetime of cyclones (in days) that pass the central North Atlantic (50°-10° W, 40°-60°N) at any time during their existence. For the hindcasts individual cyclone tracks of all 5 members and lead winters 2-5 of the initializations 1978-2012 are used, for ERA-Interim individual cyclone tracks of the period 1979/80-2016/17 are used. Vertical dashed lines denote the 0.75 quantile, i.e. the 25% strongest / longest-lived cyclones of each sample.

average extended stormtrack in HR and the related bias reductions found on the northern side and downstream end of the stormtrack. However, there is also an area of a significant reduction of the anomaly correlation for the stormtrack over Northern Canada and the Baffin Bay. Interestingly, the ~~ensemble mean~~ increase in resolution merely has an influence on the ensemble mean stormtrack bias along the North Atlantic Current (Fig. 2c)~~is merely influenced by the increase in resolution~~, and yet appears to have a strong influence on inter-annual variability and prediction skill in that region (Fig. ~~4~~5e).

The anomaly correlation between LR and ERA-Interim for the winter blocking frequencies ~~are~~ is illustrated in Fig. ~~4~~5b. Similar to the stormtrack, there is skill over Canada (r=0.3). Although the correlation is positive in large areas over the North Atlantic and Central Europe (r=0.2), it is only significant at a few of those grid points, e.g. south of Iceland and around the Baltic Sea. This changes in the HR system, where larger areas around and downstream of Newfoundland (r=0.4) and over Northern and Eastern Europe (r=0.3) show skill for the winters 2-5 (Fig. ~~4~~5d). Also, ~~some~~ large areas of significantly negative correlation over the central North Atlantic around 40°N, the Mediterranean and Scandinavia, present in LR, are reduced in size or converted to positive correlation in HR. A significant improvement in correlation with respect to the blocking frequency is therefore found for several areas, such as east of Newfoundland and all around Europe, i.e. the Mediterranean, Eastern and Northern Europe (Fig. ~~4~~5f) - except for Central Europe, which actually suffers from a significant decrease in correlation from LR to HR. The concurrence of the skill improvement around the Mediterranean and downstream of Newfoundland ~~matches well with~~ and the bias reduction in ~~those areas. The change in the anomaly correlation of the other regions cannot be directly explained by climatological changes~~the same areas again speaks for an overall better representation of the blocking dynamics in HR.

For ~~winter~~ cyclone frequencies in the winters 2-5 in LR (Fig. 6a) there is a small area of significant skill over the Arctic Ocean north of Scandinavia (r=0.2), however the rest of the domain is dominated by ~~small~~ even smaller or negative correlation (~~Fig. 5a~~r=-0.3 to r=0.2). There are large regions with significantly negative correlation west of Great Britain (r=-0.3) and over the Mediterranean (r=-0.4). Once again, with the increase in resolution ~~in~~ the skill strongly improves. In HR (Fig. ~~5c) this strongly improves and~~ 6c) positive anomaly correlation bestrides the entire North Atlantic, and the prediction is skillful (significant correlation) over a large contiguous area over the North Sea and Scandinavia (r=0.4) and at scattered grid points over the central North Atlantic (r=0.3). Only a small area over the Hudson Strait shows significant negative correlation in HR (r=0.3). Thus, the skill for extra-tropical cyclone frequencies is significantly improved through the finer resolution in large areas over the central and eastern North Atlantic, the North Sea, Scandinavia and Eastern Europe (Fig. ~~5~~6e). Those areas in which the skill is improved in HR coincide with the location of the maximum bias improvement and with the more accurately represented climatological cyclone frequencies on the downstream end along the European west coast. The analysis also reveals that not only the skill for the cyclone track frequency improves, but also for the cyclone genesis frequency, i.e. the location where the cyclones form (not shown). There is significant skill improvement from LR to HR of the cyclogenesis frequency south of Greenland, over the entire eastern North Atlantic and over Northern Europe (not shown), indicating that not only the lifetime and pathway of existing maritime cyclones is improved but also

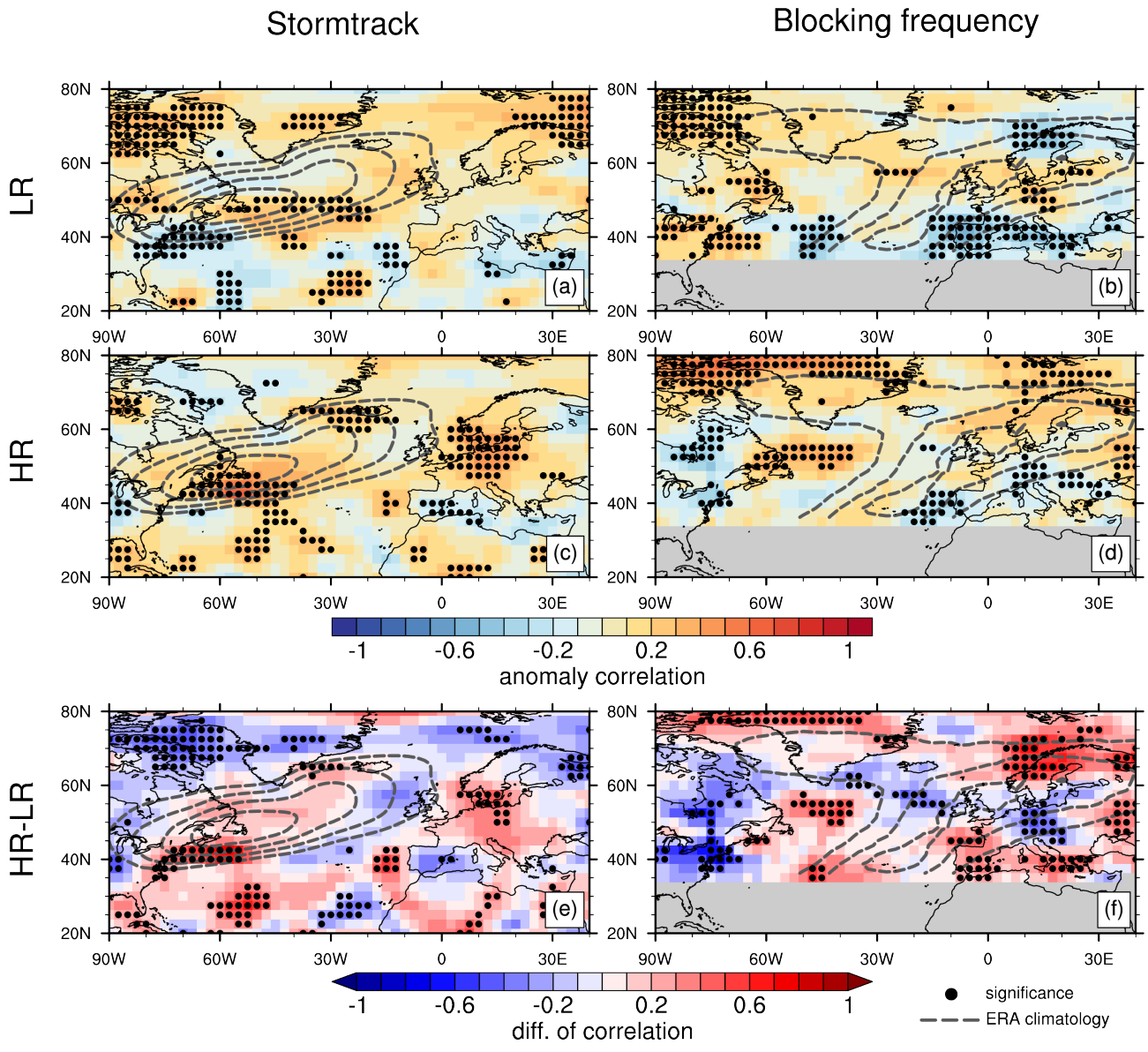
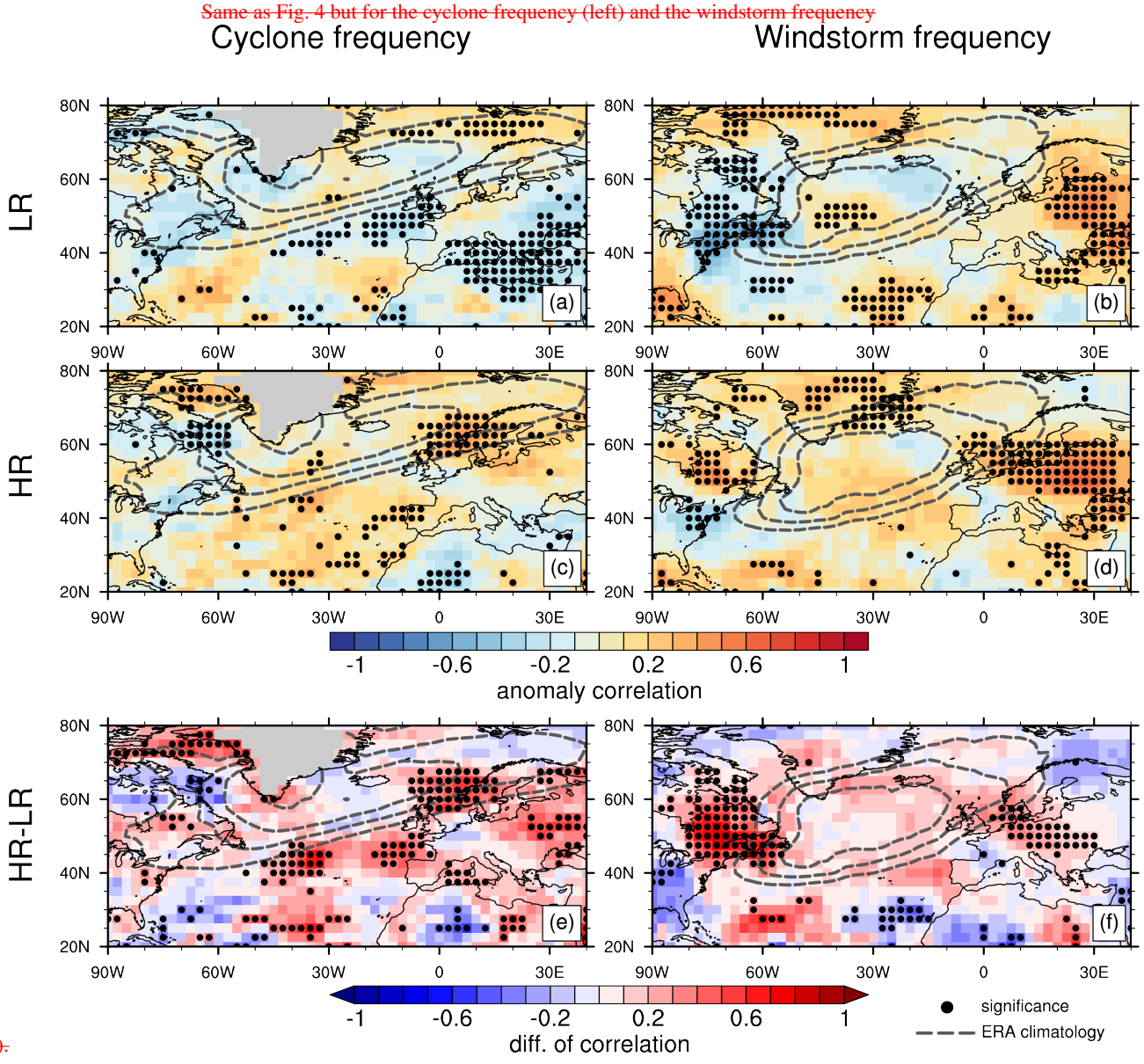


Figure 5. Anomaly correlation between the respective circulation quantity in ERA-Interim and LR (top row), between ERA-Interim and HR (middle row); and the difference between middle and top row (bottom row). The circulation quantities displayed are the stormtrack (left) and the blocking frequency (right). Initializations from the period 1978-2012 are used for both LR and HR and the correlation is computed for the winter (Oct-Mar) average of the hindcast winters 2-5. The dots mark significance (1000 times resampling of reanalysis-hindcast time series) at the 95% significance level. The dashed contours show the climatology of the circulation quantity in ERA-Interim (1979/80-2016/17) - as depicted in Fig. 1.~~The dots mark significance (1000 times resampling of reanalysis-hindcast time series).~~

the genesis of cyclones that form just off the European west coast and continental cyclones.

Prediction skill for the winter windstorms in the LR prediction system is present over the central North Atlantic ~~,~~ (r=0.2) in the region of the maximum of the windstorm climatology ~~,~~and over Eastern Europe (~~Fig. 5~~r=0.5; Fig. 6b). A large area of significant negative anomaly correlation is located around Newfoundland (r=0.6). It is remarkable that with the finer resolution the skill increases almost throughout the entire domain, i.e. it improves over the ocean but also and most strongly over continental areas (Fig. 6f). This effect is strongest and significant around Newfoundland and over Central and Eastern Europe~~(Fig. 5f)~~. This matches the results for the skill improvement of the cyclone frequencies in Fig. ~~5~~6e, indicating that if the cyclone tracks are improved along the European west coast, the downstream impact of the associated wind fields of strong cyclones is also im-

~~Same as Fig. 4 but for the cyclone frequency (left) and the windstorm frequency~~



~~(right).~~

Figure 6. Same as Fig. 5 but for the cyclone frequency (left) and the windstorm frequency (right).

proved. Also, the skill improvement over Canada and Newfoundland ~~is in line~~ again coincides with the bias reduction of the ensemble mean windstorm climatology in this region. The HR system thus produces skillful windstorm predictions over large regions of the Northern Hemisphere, e.g. eastern Canada (r=0.5), but most impressively over Central and Eastern Europe (~~Fig. 5~~r=0.6; Fig. 6d).

~~This study has evaluated the changes in the deterministic decadal forecast skill of the atmospheric extra-tropical winter circulation in response to an increase in the horizontal and vertical resolution of the forecast system , under otherwise unchanged conditions (initialization technique, parametrization). Two hindcast sets initialized in the period 1978-2012, performed with the MiKlip pre-operational decadal prediction system, one of lower resolution(LR, atm. ~1.8°, ocean ~1.5°) and one of higher resolution (HR, atm.~0.9°, ocean ~0.4°), have been evaluated for the winters 2-5 after the initialization, using 5 members each. The forecast skill has been analyzed in terms of anomaly correlation for the stormtrack, blocking frequency, cyclone frequency and windstorm frequency. Additionally, the analysis of the ensemble mean model bias has provided additional insights into the modified atmospheric dynamics and into possible sources of improved forecast skill in the higher resolved system.~~

It has been demonstrated that with the increase in the horizontal and vertical resolution, the representation of the mid-latitude dynamics in the MPI-ESM decadal prediction system is significantly improved. This applies to the ensemble mean climatology as well as the decadal prediction skill.

The stormtrack climatology in LR is represented too zonally and slightly shifted southward compared to the reanalysis, which is a well known weakness of many climate models and was discussed for

4 Discussion

The LR system shows bias patterns that are quite common in climate models of moderate resolution: a North Atlantic stormtrack which is oriented rather zonally and is southward displaced (as found e.g. for CMIP5 models models in Zappa et al. 2013) and accordingly too low blocking frequencies over the eastern North Atlantic and Northern Europe (as documented e.g. in Zappa et al. (2013). The increased model resolution counteracts this bias, leading to an extended and slightly more tilted stormtrack, but it cannot fully compensate for it. Changes are strongest on the northward side of the stormtrack maximum. These results correspond by Scaife et al. 2011). With respect to the stormtrack, the too zonal pattern in LR is to some extent corrected in HR, especially over Northern Europe. This corresponds to findings from Müller et al. (2018), who noted note a reduced bias of the atmospheric jet stream position in the northern extra-tropicsand , a decrease of the storm track bias over the northern North Atlantic in HR.

The slightly better representation of the stormtrack over the eastern North Atlantic is in line with a slightly improved blocking frequency, this means less dominance of westerlies along the European west coast and instead more influence by blocking situations. However, amongst all variables, and an increased storm activity over northern Europe for the uninitialized runs of the blocking frequency is affected least by the increased resolution. The lack of blocking over the eastern higher resolution system. Thus, in HR the stormtrack reaches further across the North Atlantic and the southward shift of the blocking climatology in LRis only marginally modified by is more tilted towards Northern Europe and to that effect is closer to the observed stormtrack than in LR. This is conform with the results presented by Zappa et al. (2013) who state that only the higher resolution - effects are strongest over the Mediterranean and downstream the southern tip of Greenland. This bias pattern is common in climate models, and has been reported e.g. by Scaife et al. (2010) and also for the MPI-ESM models of their study are able to capture the tilt of the winter stormtrack. However, the general bias of a southward displacement over the central North Atlantic is still apparent in HR, in agreement with the too zonal North Atlantic Current identified in both model versions by Müller et al. (2018).

Cyclone frequencies are overall too high in the entire domain in the LR version - a feature that has been found in previous studies with the MPI-ESM (Kruschke et al., 2014) and its predecessor ECHAM5 (Bengtsson et al., 2006). This cyclone bias is strongest over the The corresponding deficit in blocking over the Atlantic, which is strongest near the end of the stormtrack in both systems, has also been reported in previous works (e.g. by Scaife et al., 2010; **?**). While other studies that investigate the effect of the resolution on the blocking frequency bias find an increase over central and eastern Europe (**?**) or northern Europe (**?**), HR shows the increase of blocking frequency over the central and eastern North Atlantic just west of Great Britain and is consistent with a negative sea-level pressure bias , i.e. systematically too low pressure values between 2 and 5 hPa, in - as seen in Müller et al. (2018). However the changes in our study are small and a widespread negative bias remains along the European coast. A comparable result is documented in **?**, who state that the underestimation of blocking frequencies over the North Atlantic found in their lower resolution (1°) model version mainly persits in the higher resolution (0.25°) version. They conclude that, in contrast to North Pacific blocking, North Atlantic blocking is mainly driven by low-frequency eddies, which are not influenced by the same region found in LR (Müller et al., 2018). Earlier, we argued thatstrong cyclones are usually accompanied by windstorms and that because of the different bias patterns of these two variables, higher resolution.

In contrast to the previous discussed quantities, cyclone frequencies are affected very strongly by the increase in resolution. The intense positive cyclone frequency bias, which is visible over the entire North Atlantic in LR, is almost entirely removed in HR. A similar bias tendency and pattern, however not as strong, is reported in previous studies with MPI-ESM predecessors (Kruschke et al., 2014; Bengtsson et al., 2006). First analyses reveal that the intense strong bias seen in the LR cyclones is likely due to weak and moderate systems. This is in line with the evaluation of Kruschke et al. (2014), who actually showed that in LR the strong positive cyclone bias can mainly be attributed to weak and moderate systems, by illustrating a remarkably reduced bias over the North Atlantic and Europe when only intense cyclones, i.e. the strongest 25% in terms of the Laplacian of the sea-level pressure, are considered. In contrast to the minor influence on blocking, in our study appears to be the result of a combined effect of the LR system and the increase in model resolution has a powerful effect on cyclone frequencies and successfully manages to decrease the strong Atlantic initialization, as there is no such bias present in the respective uninitialized simulations of the LR system (not shown) nor in the initialized HR system. In fact, we find that this strong

cyclone frequency bias is, in the same order of magnitude, already inherent in the initialized runs of the previous MPI-ESM-LR version termed Baseline 1, analyzed by Kruschke et al. (2014) - but they only show biases for the uninitialized simulations which do not exhibit this cyclone frequency bias~~to a minimum, leaving its climatology to resemble that of the reanalysis very closely. This is also in line with the reduced sea-level pressure bias in HR found by Müller et al. (2018)~~, neither in LR nor in HR (not shown). The proportion of the bias of a) the different reanalysis used in Kruschke et al. 2014 and b) the model physics, that has since been further developed, is negligible (not shown). Nevertheless, we would like to emphasize that there is no such cyclone frequency bias in the HR system, which makes it suitable for studies on the variability and decadal prediction skill of extra-tropical cyclones.

In line with the southward shifted exit region of the stormtrack in LR there is a positive windstorm frequency bias over Central Europe and the Mediterranean. Also, the central North Atlantic is experiencing an underestimation of windstorms. A similar bias pattern is identified by Kruschke et al. (2016) for the uninitialized runs of a previous MPI-ESM system and by Befort et al. (2019) for the ECMWF's seasonal forecast systems S3 and S4. This bias cannot be corrected by the higher resolution, neither in our study nor in the higher resolved S4 system in Befort et al. (2019), and the ~~low resolution system the windstorm frequency is too small over the Atlantic Ocean and too high over land, a phenomenon that has been reported previously, e. g. by Kruschke et al. (2016). The higher resolutionseems to improve especially the strong windstorm biases along North American coastlines, i.e. the Hudson Bay, Newfoundland and Nova Scotia. Although the tilt of the windstorm track density over the North Atlantic is mended,~~ bias over the Mediterranean in HR is even aggravated. On the other hand, a slight bias improvement is found along the North Atlantic current and a strong improvement over Canada. A better representation of near surface processes would seem to be a likely cause, as the windstorms are identified from low level wind speed, but neither sea surface or 2m temperatures nor sea ice fraction show a considerable improvement over that area in the higher resolution system (Müller et al., 2018). Only the sea-level pressure shows a bias reduction over Canada, however the opposite is the ~~model still generates too many windstorms over Europe and the positive bias there is generally amplified.~~case over the North Atlantic Current.

~~With respect to the decadal prediction skill this analysis showed that the increased resolution of the MPI-ESM decadal prediction system significantly improves~~ Although the bias and the anomaly correlation ~~in crucial regions of the North Atlantic and Europe on lead times of 2-5 winters for all four extra-tropical circulationmeasures. Furthermore, the areas with improved forecast skill are key regions for the genesis of synoptic weather systems in the North Atlantic and for their impact on Europe. This is particularly evident for the stormtrack, for which~~ are per se unrelated, they are both important metrics to assess the model's performance to correctly represent the mean state and the variability of the atmospheric circulation. If they appear to be improved in the same location this does not imply a causal interrelation. However, it all the more indicates that local physical processes are improved in the higher resolution prediction system.

With respect to the decadal prediction skill, for the stormtrack a strong and significant skill improvement is found along the North Atlantic Current and over Central Europe. Given the important role of surface heat fluxes and local SST gradients for the dynamics of the stormtrack (Brayshaw et al., 2011), these are likely sources of improved atmospheric variability in the HR system. The skill improvement over Central Europe is in line with the bias improvement of the stormtrack at its downstream end (stronger tilt and downstream extension~~of the stormtrack climatology in HR results in improved and significant decadal forecast skill east of Greenland and in Central Europe.~~) shown in our study as well as with reduced sea surface temperature and salinity biases over the eastern and northern North Atlantic found in Müller et al. (2018) for the uninitialized runs of the same model system. However, the influence of local SST gradients on the stormtrack skill improvement along the North Atlantic current is debatable, given the mostly unchanged bias of the North Atlantic Current in HR documented in Müller et al. (2018).

~~Significant~~ The skill for blocking frequencies over the North Atlantic and European domain is basically not existent in the LR system, it shows only skill over Canada. A similar pattern of skill is found in **?** for seasonal forecasts performed with the CCMC model, which is as well based on the ECHAM. They further find that this lower resolution model (CCMC; atm. ~1.875°) underrepresents the variance of blocking in the eastern North Atlantic more than the higher resolution model (UKMO; atm. 0.83° x 0.55°) of their study. We find in HR a significant improvement in the anomaly correlation of the blocking frequency ~~is found~~ downstream of where the stormtrack skill is improved, ~~and in large patches all around Europe, except for Central Europe . The strongest effect of the bias reduction , found over the Mediterranean, coincides with skill improvement in that area, however since the bias reduction is generally marginal, a direct effect on the decadal prediction skill is not necessarily given~~i.e. over the central North Atlantic and Northern Europe and the Mediterranean. The latter coincides with a strong bias reduction in that area. This matches with the results of **?**, where as well only the higher resolution system shows skill for blocking frequencies in the Euro-Atlantic sector - in their case over the eastern North Atlantic and Central Europe. They state that those are preliminary the regions where blocking activity is strong and related to the NAO variability. If this relation was as well valid for our simulations, the

skill improvement for blocking over Northern Europe and the Mediterranean would be in line with NAO amplitudes reaching further towards Europe in HR than in LR, as found by (Müller et al., 2018).

The strong misrepresentation of the cyclone ~~climatology~~ frequencies in LR results in no decadal forecast skill throughout the North Atlantic and European domain. ~~Thus~~ But with the increase in resolution not only the striking climatological bias reduction ~~in HR also impacts~~ is achieved, but also the prediction skill ~~, which~~ is improved throughout the entire domain ~~(significantly along the outskirts of the cyclone frequency maximum) and results~~ - significantly over the central and eastern North Atlantic and Northern Europe - resulting in skillful cyclone frequency predictions over Northern Europe in HR. The improved representation of cyclones in this region may also be beneficial for the prediction of blocking over Scandinavia (where the skill in HR is significantly improved), as cyclones can contribute to downstream blocking formation through eddy vorticity forcing (Shutts, 1983) and diabatic processes (Pfahl et al., 2015). ~~A~~ Apart from the removed initialization effect in HR, a more accurate representation of smaller-scale diabatic processes may ~~also~~ be a reason for the increased forecast skill of cyclones at the southern flank of the main stormtrack, over the subtropical North Atlantic and the Mediterranean, as moist processes are thought to be particularly important for such subtropical systems (e.g. Davis, 2010). The fact, that Kruschke et al. (2014) find the decadal forecast skill to be higher for strong cyclones than for all cyclones in the former LR system could as well be related to the initialization effect apparent in LR. As strong cyclones are not affected by this initialization induced bias, likely their prediction skill is as well more credible.

~~In~~ Although the forecast skill for 10m wind speeds and wind energy output only differs slightly between different ocean initializations (Moemken et al., 2016), this study reveals that an increased model resolution has a large impact on the hindcast skill of synoptic scale features, such as cyclones and windstorms. Thus, in line with these skill improvements in the cyclone frequency, the skill for windstorms ~~improves as well~~ also improves significantly over North-East and Central Europe, i.e. south of the cyclones' signal. This matches with the general south-eastward displacement of the maximum wind speeds relative to the cyclone center (Leckebusch et al., 2008). Müller et al. (2018) deduced from the stormtrack bias changes they found in the uninitialized higher resolution runs that more storms entering the northern European region can be expected, relative to LR. Although this cannot be confirmed with respect to the windstorm frequency climatology in our study, its prediction skill is significantly improved from the North Sea through Eastern Europe. Different studies suggest that a better representation of the North Atlantic current in the model would contribute to a better representation of the storm track (e.g. Brayshaw et al., 2011; Scaife et al., 2011) and thus would probably lead to increased downstream predictive skill for cyclones and windstorms (Kruschke et al., 2014). Although the improvement for the North Atlantic current in terms of sea surface temperature and salinity is small for the HR system, as reported by Müller et al. (2018) for the uninitialized runs, this study found improved decadal prediction skill downstream of the stormtrack, not only for cyclones and windstorms but also for blocking frequencies. This indicates that the variability of strong North Atlantic cyclones ~~traveling towards Scandinavia and leading to~~ causing windstorms in North and Central Europe is much better captured by ~~the high resolution decadal prediction system. Interestingly, the skill is not negatively affected in South-East Europe, although the climatological windstorm bias is amplified in HRin that region. On the other hand, the strong bias reduction over North America and Canada appears to impact the prediction skill, thus a significant skill improvement for windstorm frequency is found over Canada and parts of~~ HR.

5 Conclusions

This study evaluated the response of the deterministic decadal forecast skill of the atmospheric extra-tropical winter circulation to an increase in the resolution of the forecast system. This was performed under otherwise unchanged conditions, i.e. the same numerical model, initialization technique and parametrization were used and only the resolution of the model was changed. The two hindcast sets (LR: atm. ∼1.8°, ocean ∼1.5° and HR: atm. ∼0.9°, ocean ∼0.4°) were initialized in the period 1978-2012 and evaluated for the winters 2-5 after the initialization, using 5 members each. Those hindcasts were performed with the MiKlip pre-operational decadal prediction system, based on the MPI-ESM. The forecast skill was analyzed over the North Atlantic ~~Current. Although, the forecast skill for 10m wind speeds and wind energy output only differs slightly between different ocean initializations (Moemken et al., 2016), this study reveals that the increased resolution has a large impact on the hindcast skill of synoptic scale features, such as cyclones and windstorms.~~ region in terms of anomaly correlation for the stormtrack, blocking frequency, cyclone frequency and windstorm frequency. ERA-Interim, i.e. the winterly averages of the four quantities between 1979/80 and 2016/17, served as the reference dataset. The analysis of the ensemble mean model bias has provided additional insights into the modified atmospheric dynamics and into possible sources of improved forecast skill in the higher resolved system.

~~Overall we demonstrated that there is a chain of~~ In summary we demonstrated an improvement of the mid-latitude dynamics in the North Atlantic region with

an increase in the model resolution. This comprises an improvement of both the mean state (climatology) and the temporal variability (decadal prediction skill~~improvement amongst the~~) for the different extra-tropical circulation metrics~~with the increase in model resolution, similar to the interrelations laid out in Scaife et al. (2011). Also~~. Although there are yet no other studies on this topic with respect to decadal time scales, our results are in agreement with ~~previous studies by Prodhomme et al. (2016) and Befort et al. (2019) who found skill improvements in different seasonal prediction systems~~ findings from seasonal prediction studies (e.g. Prodhomme et al., 2016; Befort et al., 2019), who showed skill improvements for blocking, windstorm and cyclone frequencies when the ~~model~~ same model is used and only the resolution is increased. ~~Our study showed that there is~~

The improvements found in our study for the different metrics follow a physically consistent line of argument and the areas of improved forecast skill are crucial regions for the genesis and intensification of synoptic weather systems over the North Atlantic and for their impact on Europe. Thus, we identified a significant improvement of the stormtrack skill along the North Atlantic Current ~~followed by~~ (i.e. the source region of synoptic eddies), a downstream improvement of the cyclone frequency skill over the central North Atlantic (where the synoptic systems intensify), and finally improved skill of the cyclone, windstorm and blocking frequencies over the ~~impact area Europe~~European continent (i.e. the impact area). Additionally, not only does the prediction skill improve with a finer ~~grid~~ resolution (HR vs. LR), the HR system itself offers significant deterministic ~~forecast skill~~ decadal forecast skill for the extra-tropical circulation metrics in large regions over the North Atlantic and Europe (HR vs. ERA-Interim) ~~. An important question remains, as to which physical processes form the basis of this detected~~ for the considered lead time of 2-5 winters.

By analyzing different but linked extra-tropical circulation metrics, this study contributes to the elucidation of the processes that lead to the decadal prediction skill ~~of the different circulation variables, and should be explored in future research.~~ in the North Atlantic region. Our results are encouraging as they document the successful advancement of decadal prediction systems and in particular of the deterministic decadal prediction skill of extra-tropical features and extreme events. However, future studies using different prediction systems, possibly of higher resolution and larger ensemble sizes, and especially rather process oriented analyses will be needed to shed further light on the robustness of our results and the sources of the presented skill.

Competing interests. No competing interests are present.

Acknowledgements. The authors would like to acknowledge funding from CoreLogic SARL Paris and from the Federal Ministry of Education and Research in Germany (BMBF) through the research program MiKlip II (FKZ: 01LP1519A, 01LP1519B, 01LP1520A) and the CMIP6-DICAD project (FKZ: 01LP1605D). We acknowledge support by the Open Access Publication Fund of the Freie Universität Berlin. We would also like to thank the anonymous reviewers for their valuable comments.

References

Balmaseda, M. A., Mogensen, K., and Weaver, A. T.: Evaluation of the ECMWF ocean reanalysis system ORAS4, Quarterly Journal of the Royal Meteorological Society, 139, 1132–1161, https://doi.org/10.1002/qj.2063, 2013.

Barnes, E. A., Slingo, J., and Woollings, T.: A methodology for the comparison of blocking climatologies across indices, models and climate scenarios, CLIMATE DYNAMICS, 38, 2467–2481, https://doi.org/10.1007/s00382-011-1243-6, 2012.

Befort, D. J., Wild, S., Knight, J. R., Lockwood, J. F., Thornton, H. E., Hermanson, L., Bett, P. E., Weisheimer, A., and Leckebusch, G. C.: Seasonal forecast skill for extratropical cyclones and windstorms, Quarterly Journal of the Royal Meteorological Society, 145, 92–104, https://doi.org/10.1002/qj.3406, 2019.

Bengtsson, L., Hodges, K. I., and Roeckner, E.: Storm tracks and climate change, Journal of Climate, 19, 3518–3543, https://doi.org/10.1175/JCLI3815.1, 2006.

Blackmon, M. L., Wallace, J., Lau, N., and Mullen, S. L.: An observational study of the Northern Hemisphere wintertime circulation, Journal of Atmospheric Sciences, 34, 1040–1053, https://doi.org/10.1175/1520-0469(1977)034<1040:AOSOTN>2.0.CO;2, 1976.

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., Kushnir, Y., Kimoto, M., Meehl, G. A., Msadek, R., Müller, W. A., Taylor, K. E., Zwiers, F., Rixen, M., Ruprich-Robert, Y., and Eade, R.: The Decadal Climate Prediction Project (DCPP) contribution to CMIP6, GEOSCIENTIFIC MODEL DEVELOPMENT, 9, 3751–3777, https://doi.org/10.5194/gmd-9-3751-2016, 2016.

Brayshaw, D. J., Hoskins, B., and Blackburn, M.: The Basic Ingredients of the North Atlantic Storm Track. Part II: Sea Surface Temperatures, Journal of the Atmospheric Sciences, 68, 1784–1805, https://doi.org/10.1175/2011JAS3674.1, https://doi.org/10.1175/2011JAS3674.1, 2011.

Davis, C. A.: Simulations of Subtropical Cyclones in a Baroclinic Channel Model, Journal of the Atmospheric Sciences, 67, 2871–2892, https://doi.org/10.1175/2010JAS3411.1, 2010.

Dawson, A., Matthews, A. J., Stevens, D. P., Roberts, M. J., and Vidale, P. L.: Importance of oceanic resolution and mean state on the extra-tropical response to El Niño in a matrix of coupled models, Climate Dynamics, 41, 1439–1452, https://doi.org/10.1007/s00382-012-1518-6, https://doi.org/10.1007/s00382-012-1518-6, 2013.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Holm, E. V., Isaksen, L., Kallberg, P., Koehler, M., Matricardi, M., McNally,

A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

Doblas-Reyes, F. J., Andreu-Burillo, I., Chikamoto, Y., Garcia-Serrano, J., Guemas, V., Kimoto, M., Mochizuki, T., Rodrigues, L. R. L., and van Oldenborgh, G. J.: Initialized near-term regional climate change prediction, Nature Communications, 4, 1715, https://doi.org/10.1038/ncomms2704, 2013.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Boettinger, M., Brovkin, V., Crueger, T., Esch, M., Fieg, K., Glushak, K., Gayler, V., Haak, H., Hollweg, H.-D., Ilyina, T., Kinne, S., Kornblueh, L., Matei, D., Mauritsen, T., Mikolajewicz, U., Müller, W., Notz, D., Pithan, F., Raddatz, T., Rast, S., Redler, R., Roeckner, E., Schmidt, H., Schnur, R., Segschneider, J., Six, K. D., Stockhause, M., Timmreck, C., Wegner, J., Widmann, H., Wieners, K.-H., Claussen, M., Marotzke, J., and Stevens, B.: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5, Journal of Advances In Modeling Earth Systems, 5, 572–597, https://doi.org/10.1002/jame.20038, 2013.

Goddard, L., Kumar, A., Solomon, A., Smith, D., Boer, G., Gonzalez, P., Kharin, V., Merryfield, W., Deser, C., Mason, S. J., Kirtman, B. P., Msadek, R., Sutton, R., Hawkins, E., Fricker, T., Hegerl, G., Ferro, C. A. T., Stephenson, D. B., Meehl, G. A., Stockdale, T., Burgman, R., Greene, A. M., Kushnir, Y., Newman, M., Carton, J., Fukumori, I., and Delworth, T.: A verification framework for interannual-to-decadal predictions experiments, Climate Dynamics, 40, 245–272, https://doi.org/10.1007/s00382-012-1481-2, 2013.

Haas, R., Reyers, M., and Pinto, J. G.: Decadal predictability of regional-scale peak winds over Europe using the Earth System Model of the Max-Planck-Institute for Meteorology, Meteorologische Zeitschrift, 25, 739–752, https://doi.org/10.1127/metz/2015/0583, http://dx.doi.org/10.1127/metz/2015/0583, 2016.

Illing, S., Kadow, C., Oliver, K., and Cubasch, U.: MurCSS: A Tool for Standardized Evaluation of Decadal Hindcast Systems, Journal of Open Research Software, 2, 2014.

Jung, T., Miller, M. J., Palmer, T. N., Towers, P., Wedi, N., Achuthavarier, D., Adams, J. M., Altshuler, E. L., Cash, B. A., Kinter, J. L., Marx, L., Stan, C., and Hodges, K. I.: High-Resolution Global Climate Simulations with the ECMWF Model in Project Athena: Experimental Design, Model Climate, and Seasonal Forecast Skill, Journal of Climate, 25, 3155–3172, https://doi.org/10.1175/JCLI-D-11-00265.1, https://doi.org/10.1175/JCLI-D-11-00265.1, 2012.

Kadow, C., Illing, S., Kunst, O., Schartner, T., Grieger, J., Schuster, M., Richling, A., Kirchner, I., Rust, H., Cubasch, U., and Ulbrich, U.: Freva - Free Evaluation System Framework for Earth System Modeling, Geoscientific Model Development, in preparation.

Kadow, C., Illing, S., Kunst, O., Rust, H. W., Pohlmann, H., Müller, W. A., and Cubasch, U.: Evaluation of forecasts by accuracy and spread in the MiKlip decadal climate prediction system, METEOROLOGISCHE ZEITSCHRIFT, 25, 631–643, https://doi.org/10.1127/metz/2015/0639, 2016.

Kadow, C., Illing, S., Kroener, I., Ulbrich, U., and Cubasch, U.: Decadal climate predictions improved by ocean ensemble dispersion filtering, JOURNAL OF ADVANCES IN MODELING EARTH SYSTEMS, 9, 1138–1149, https://doi.org/10.1002/2016MS000787, 2017.

Kaspar, F., Rust, H. W., Ulbrich, U., and Becker, P.: Verification and process oriented validation of the MiKlip decadal prediction system, Meteorologische Zeitschrift, 25, 629–630, https://doi.org/10.1127/metz/2016/0831, http://dx.doi.org/10.1127/metz/2016/0831, 2016.

Keenlyside, N. S., Latif, M., Jungclaus, J., Kornblueh, L., and Roeckner, E.: Advancing decadal-scale climate prediction in the North Atlantic sector, NATURE, 453, 84–88, https://doi.org/10.1038/nature06921, 2008.

Kim, H. M., Webster, P. J., and Curry, J. A.: Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts, Geophysical Research Letters, 39, L10 701, https://doi.org/10.1029/2012GL051644, 2012.

Kirtman, B., Power, S., Adedoyin, J., Boer, G., Bojariu, R., Camilloni, I., Doblas-Reyes, F., Fiore, A., Kimoto, M., Meehl, G., Prather, M., Sarr, A., Schär, C., Sutton, R., van Oldenborgh, G., Vecchi, G., and Wang, H.: Near-term Climate Change: Projections and Predictability, book section 11, p. 953–1028, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/CBO9781107415324.023, www.climatechange2013.org, 2013.

Kröger, J., Pohlmann, H., Sienz, F., Marotzke, J., Baehr, J., Köhl, A., Modali, K., Polkova, I., Stammer, D., Vamborg, F. S. E., and Müller, W. A.: Full-field initialized decadal predictions with the MPI earth system model: an initial shock in the North Atlantic, Climate Dynamics, https://doi.org/10.1007/s00382-017-4030-1, 2017.

Kruschke, T.: Winter wind storms: Identication, verication of decadal predictions, and regionalization, Ph.D. thesis, Freie Universität Berlin, 2014.

Kruschke, T., Rust, H. W., Kadow, C., Leckebusch, G. C., and Ulbrich, U.: Evaluating decadal predictions of northern hemispheric cyclone frequencies, Tellus Series A-dynamic Meteorology and Oceanography, 66, 22 830, https://doi.org/10.3402/tellusa.v66.22830, 2014.

Kruschke, T., Rust, H. W., Kadow, C., Müller, W. A., Pohlmann, H., Leckebusch, G. C., and Ulbrich, U.: Probabilistic evaluation of decadal prediction skill regarding Northern Hemisphere winter storms, METEOROLOGISCHE ZEITSCHRIFT, 25, 721–738, https://doi.org/10.1127/metz/2015/0641, 2016.

Kushnir, Y., Scaife, A. A., Arritt, R., Balsamo, G., Boer, G., Doblas-Reyes, F., Hawkins, E., Kimoto, M., Kolli, R. K., Kumar, A., Matei, D., Matthes, K., Müller, W. A., O'Kane, T., Perlwitz, J., Power, S., Raphael, M., Shimpo, A., Smith, D., Tuma, M., and Wu, B.: Towards operational predictions of the near-term climate, Nature Climate Change, 9, 94–101, https://doi.org/10.1038/s41558-018-0359-7, 2019.

Leckebusch, G. C., Renggli, D., and Ulbrich, U.: Development and application of an objective storm severity measure for the Northeast Atlantic region, Meteorologische Zeitschrift, 17, 575–587, https://doi.org/10.1127/0941-2948/2008/0323, http://dx.doi.org/1fv0.1127/0941-2948/2008/0323, 2008.

Marotzke, J., Müller, W. A., Vamborg, F. S. E., Becker, P., Cubasch, U., Feldmann, H., Kaspar, F., Kottmeier, C., Marini, C., Polkova, I., Proemmel, K., Rust, H. W., Stammer, D., Ulbrich, U., Kadow, C., Koehl, A., Kroeger, J., Kruschke, T., Pinto, J. G., Pohlmann, H., Reyers, M., Schroeder, M., Sienz, F., Timmreck, C., and Ziese, M.: MIKLIP A NATIONAL RESEARCH PROJECT ON DECADAL CLIMATE PREDICTION, BULLETIN OF THE AMERICAN METEOROLOGICAL SOCIETY, 97, 2379–2394, https://doi.org/10.1175/BAMS-D-15-00184.1, 2016.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., Esch, M., Fast, I., Fiedler, S., Fläschner, D., Gayler, V., Giorgetta, M., Goll, D. S., Haak, H., Hagemann, S., Hedemann, C., Hohenegger, C., Ilyina, T., Jahns, T., Jimenéz-de-la Cuesta, D., Jungclaus, J., Kleinen, T., Kloster, S., Kracher, D., Kinne, S., Kleberg, D., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Möbis, B., Müller, W. A., Nabel, J. E. M. S., Nam, C. C. W., Notz, D., Nyawira, S.-S., Paulsen, H., Peters, K., Pincus, R., Pohlmann, H., Pongratz, J., Popp, M., Raddatz, T. J., Rast, S., Redler, R., Reick, C. H., Rohrschneider, T., Schemann, V., Schmidt, H., Schnur, R., Schulzweida, U., Six, K. D., Stein, L., Stemmler, I., Stevens, B., von Storch, J.-S., Tian, F., Voigt, A., Vrese, P., Wieners, K.-H., Wilkenskjeld, S., Winkler, A., and Roeckner, E.: Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO2, Journal of Advances in Modeling Earth Systems, 0, https://doi.org/10.1029/2018MS001400, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001400, 2019.

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., Corti, S., Danabasoglu, G., Doblas-Reyes, F., Hawkins, E., Karspeck, A., Kimoto, M., Kumar, A., Matei, D., Mignot, J., Msadek, R., Navarra, A., Pohlmann, H., Rienecker, M., Rosati, T., Schneider, E., Smith, D., Sutton, R., Teng, H., van Oldenborgh, G. J., Vecchi, G., and Yeager, S.: DECADAL CLIMATE PREDICTION An Update from the Trenches, BULLETIN OF THE AMERICAN METEOROLOGICAL SOCIETY, 95, 243–267, https://doi.org/10.1175/BAMS-D-12-00241.1, 2014.

Moemken, J., Reyers, M., Buldmann, B., and Pinto, J. G.: Decadal predictability of regional scale wind speed and wind energy potentials over Central Europe, TELLUS SERIES A-DYNAMIC METEOROLOGY AND OCEANOGRAPHY, 68, https://doi.org/10.3402/tellusa.v68.29199, 2016.

Monerie, P.-A., Coquart, L., Maisonnave, É., Moine, M.-P., Terray, L., and Valcke, S.: Decadal prediction skill using a high-resolution climate model, Climate Dynamics, 49, 3527–3550, https://doi.org/10.1007/s00382-017-3528-x, https://doi.org/10.1007/s00382-017-3528-x, 2017.

Monerie, P. A., Robson, J., Dong, B. W., and Dunstone, N.: A role of the Atlantic Ocean in predicting summer surface air temperature over North East Asia?, Climate Dynamics, 51, 473–491, https://doi.org/10.1007/s00382-017-3935-z, 2018.

Murray, R. J. and Simmonds, I.: A numerical scheme for tracking cyclone centres from digital data. Part I: Development and operation of the scheme, Australian Meteorological Magazine, 39, 155–166, 1991.

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., Bunzel, F., Esch, M., Ghosh, R., Haak, H., Ilyina, T., Kleine, T., Kornblueh, L., Li, H., Modali, K., Notz, D., Pohlmann, H., Roeckner, E., Stemmler, I., Tian, F., and Marotzke, J.: A Higher-resolution Version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR), Journal of Advances in Modeling Earth Systems, 0, https://doi.org/10.1029/2017MS001217, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2017MS001217, 2018.

Müller, W. A., Baehr, J., Haak, H., Jungclaus, J. H., Kroeger, J., Matei, D., Notz, D., Pohlmann, H., von Storch, J. S., and Marotzke, J.: Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology, GEOPHYSICAL RESEARCH LETTERS, 39, https://doi.org/10.1029/2012GL053326, 2012.

Park, T., Park, W., and Latif, M.: Correcting North Atlantic sea surface salinity biases in the Kiel Climate Model: influences on ocean circulation and Atlantic Multidecadal Variability, Climate Dynamics, 47, 2543–2560, https://doi.org/10.1007/s00382-016-2982-1, https://doi.org/10.1007/s00382-016-2982-1, 2016.

Pfahl, S., Schwierz, C., Croci-Maspoli, M., Grams, C. M., and Wernli, H.: Importance of latent heat release in ascending air streams for atmospheric blocking, Nature Geoscience, 8, 610–+, https://doi.org/10.1038/NGEO2487, 2015.

Pohlmann, H., Muller, W. A., Kulkarni, K., Kameswarrao, M., Matei, D., Vamborg, F. S. E., Kadow, C., Illing, S., and Marotzke, J.: Improved forecast skill in the tropics in the new MiKlip decadal climate predictions, Geophysical Research Letters, 40, 5798–5802, https://doi.org/10.1002/2013GL058051, 2013.

Polkova, I., Brune, S., Kadow, C., Romanova, V., Gollan, G., Baehr, J., Glowienka-Hense, R., Greatbatch, R. J., Hense, A., Illing, S., Köhl, A., Kröger, J., Müller, W. A., Pankatz, K., and Stammer, D.: Initialization and Ensemble Generation for Decadal Climate Predictions: A Comparison of Different Methods, Journal of Advances in Modeling Earth Systems, 11, 149–172, https://doi.org/10.1029/2018MS001439, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001439, 2019.

Prodhomme, C., BattÃ©, L., Massonnet, F., Davini, P., Bellprat, O., Guemas, V., and Doblas-Reyes, F. J.: Benefits of Increasing the Model Resolution for the Seasonal Forecast Quality in EC-Earth, Journal of Climate, 29, 9141–9162, https://doi.org/10.1175/JCLI-D-16-0117.1, https://doi.org/10.1175/JCLI-D-16-0117.1, 2016.

Robson, J., Polo, I., Hodson, D. L. R., Stevens, D. P., and Shaffrey, L. C.: Decadal prediction of the North Atlantic subpolar gyre in the HiGEM high-resolution climate model, Climate Dynamics, 50, 921–937, https://doi.org/10.1007/s00382-017-3649-2, 2018.

Scaife, A. A., Woollings, T., Knight, J., Martin, G., and Hinton, T.: Atmospheric Blocking and Mean Biases in Climate Models, JOURNAL OF CLIMATE, 23, 6143–6152, https://doi.org/10.1175/2010JCLI3728.1, 2010.

Scaife, A. A., Copsey, D., Gordon, C., Harris, C., Hinton, T., Keeley, S., O'Neill, A., Roberts, M., and Williams, K.: Improved Atlantic winter blocking in a climate

model, GEOPHYSICAL RESEARCH LETTERS, 38, https://doi.org/10.1029/2011GL049573, 2011.

Scherrer, S. C., Croci-Maspoli, M., Schwierz, C., and Appenzeller, C.: Two-dimensional indices of atmospheric blocking and their statistical relationship with winter climate patterns in the Euro-Atlantic region, International Journal of Climatology, 26, 233–249, https://doi.org/10.1002/joc.1250, 2006.

Shaffrey, L. C., Stevens, I., Norton, W. A., Roberts, M. J., Vidale, P. L., Harle, J. D., Jrrar, A., Stevens, D. P., Woodage, M. J., Demory, M. E., Donners, J., Clark, D. B., Clayton, A., Cole, J. W., Wilson, S. S., Connolley, W. M., Davies, T. M., Iwi, A. M., Johns, T. C., King, J. C., New, A. L., Slingo, J. M., Slingo, A., Steenman-Clark, L., and Martin, G. M.: U.K. HiGEM: The New U.K. High-Resolution Global Environment Model - Model Description and Basic Evaluation, Journal of Climate, 22, 1861–1896, https://doi.org/10.1175/2008JCLI2508.1, https://doi.org/10.1175/2008JCLI2508.1, 2009.

Shutts, G. J.: The Propagation of Eddies In Diffluent Jetstreams - Eddy Vorticity Forcing of Blocking Flow-fields, Quarterly Journal of the Royal Meteorological Society, 109, 737–761, https://doi.org/10.1002/qj.49710946204, 1983.

Smith, D. M., Cusack, S., Colman, A. W., Folland, C. K., Harris, G. R., and Murphy, J. M.: Improved surface temperature prediction for the coming decade from a global climate model, SCIENCE, 317, 796–799, https://doi.org/10.1126/science.1139540, 2007.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bulletin of the American Meteorological Society, 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, https://doi.org/10.1175/BAMS-D-11-00094.1, 2012.

Uppala, S. M., Kallberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Van De Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Holm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., McNally, A. P., Mahfouf, J. F., Morcrette, J. J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J.: The ERA-40 re-analysis, Quarterly Journal of the Royal Meteorological Society, 131, 2961–3012, https://doi.org/10.1256/qj.04.176, 2005.

Wang, C., Zhang, L., Lee, S.-K., Wu, L., and Mechoso, C. R.: A global perspective on CMIP5 climate model biases, Nature Climate Change, 4, 201, https://doi.org/10.1038/nclimate2118, 2014.

Xin, X. G., Gao, F., Wei, M., Wu, T. W., Fang, Y. J., and Zhang, J.: Decadal prediction skill of BCC-CSM1.1 climate model in East Asia, International Journal of Climatology, 38, 584–592, https://doi.org/10.1002/joc.5195, 2018.

Zappa, G., Shaffrey, L. C., and Hodges, K. I.: The Ability of CMIP5 Models to Simulate North Atlantic Extratropical Cyclones, Journal of Climate, 26, 5379–5396, https://doi.org/10.1175/JCLI-D-12-00501.1, 2013.