Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2019-18-AC2, 2019
© Author(s) 2019. This work is distributed under
the Creative Commons Attribution 4.0 License.

Earth System Dynamics
Discussions

Interactive comment on “Improvement in the decadal prediction skill of the northern hemisphere extra-tropical winter circulation through increased model resolution” *by* Mareike Schuster et al.

Mareike Schuster et al.

mareike.schuster@met.fu-berlin.de

Received and published: 10 July 2019

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)."
Added to p.14, l.23:
"A similar change in blocking frequencies with increased model resolution was also found in Davini et al. (2017)."

- 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 'Pre-operational' (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 (+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 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 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 analy-

ses. 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. 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, ie 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 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 extratropical circulation metrics. However, we now added this information also to the discussion and inserted the following to p.16 l.17:
  "To this date there is no study that addresses the effect which the model resolution has on the decadal prediction skill on extratropical circulation metrics. However, our results are in agreement. . ."

- 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 that we analyze.”

- 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.

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2019-18, 2019.

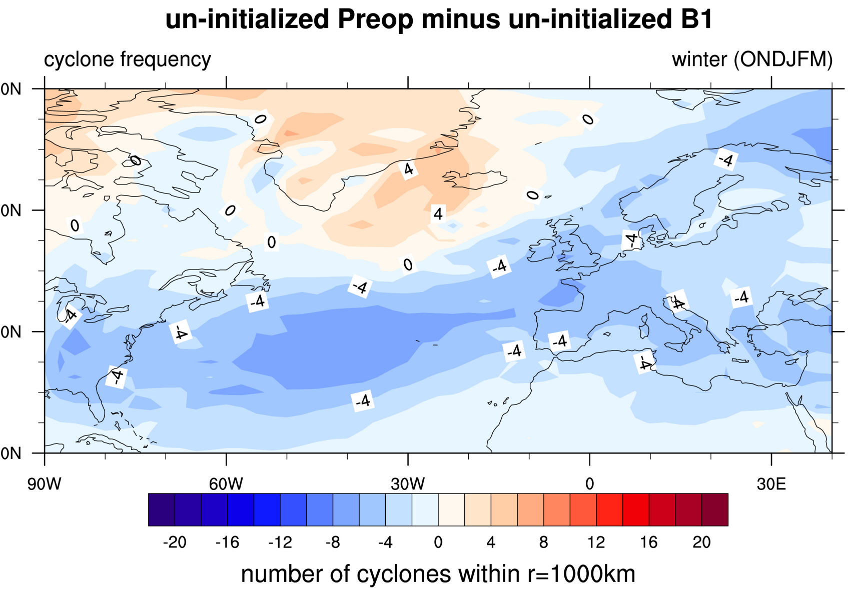

Fig. 4. Effect of the model development - Difference of the cyclone frequency between the un-initialized Preop-LR and un-initialized B1-LR simulations

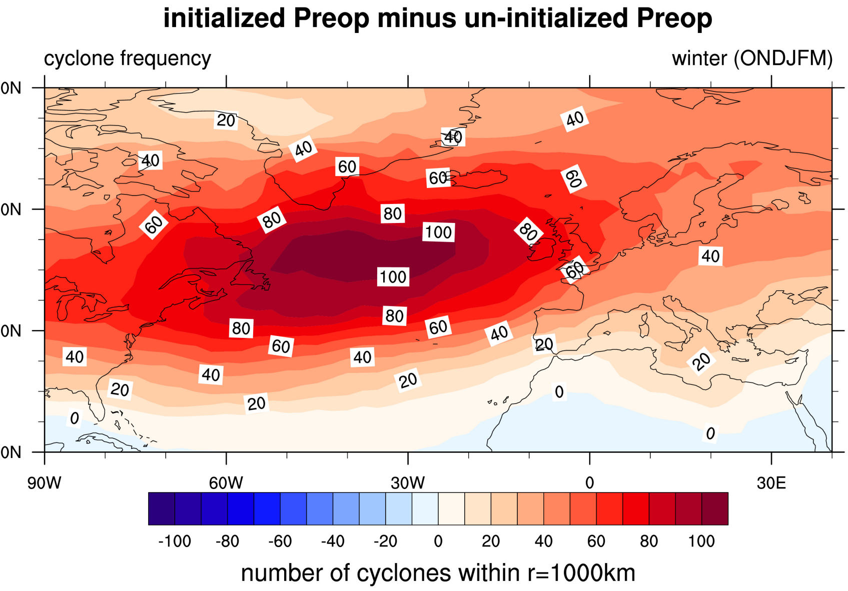

Fig. 5. Effect of the initialization in Preop-LR - Difference of the cyclone frequency between the initialized Preop-LR and un-initialized Preop-LR simulations

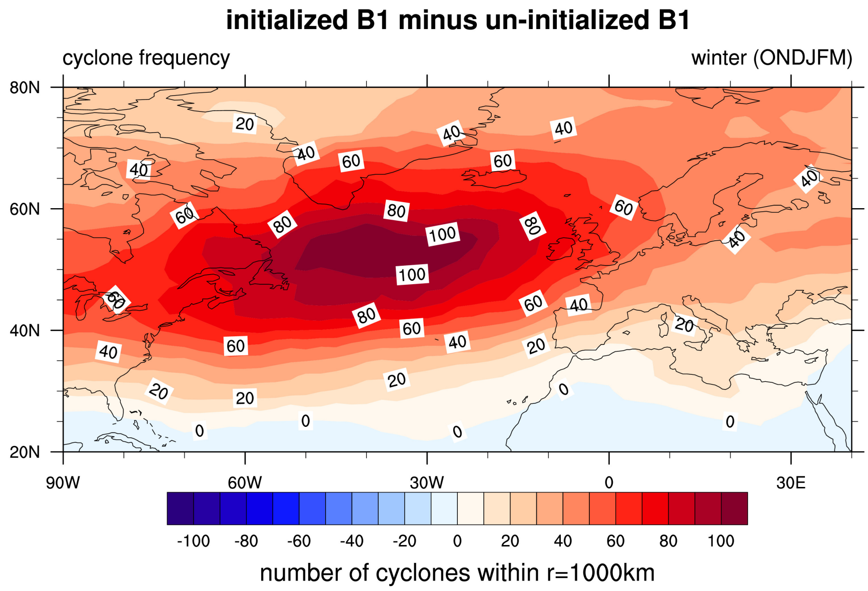

Fig. 6. Effect of the initialization in B1-LR - Difference of the cyclone frequency between the initialized B1-LR and un-initialized B1-LR simulations

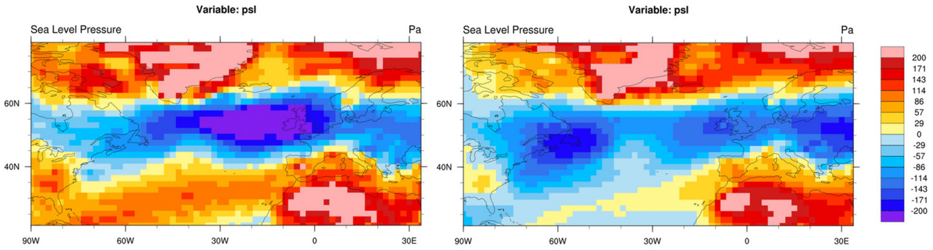

Fig. 7. Mean Sea-Level Pressure bias relative to ERA-Interim - left: in the Preop-LR system; right: in the Preop-HR system

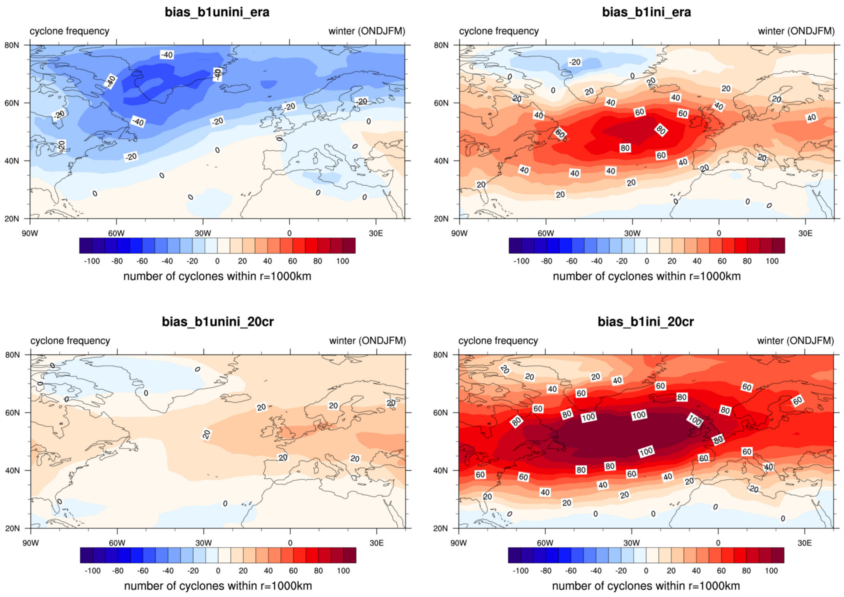

Fig. 8. Cyclone frequency bias in the different simulations relative to different reanalyses - top: relative to ERA-Interim; bottom: relative to 20CR; left: un-initialized simul.; right: initialized simul.

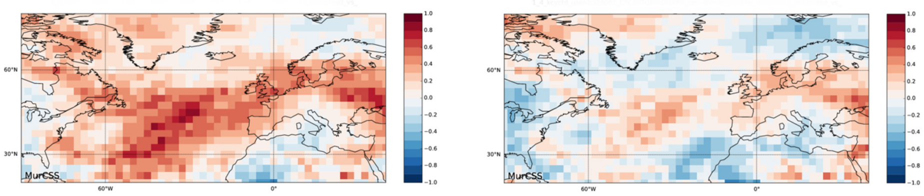

Fig. 9. Spread vs. MSE (Logarithmic Ensemble Spread Score - LESS) for the cyclone frequency - left: in the Preop-LR system; right: in the Preop-HR system