**MWR-D-22-0318, Biernat et al., Response to reviewers.**

The authors thank the reviewers for their thorough and thoughtful reviews. Our responses (red) to the comments (black) are given below. We provide line numbers corresponding to the revised manuscript to indicate where changes were made. We will first address a main concern from Reviewer 2 since this concern has resulted in us only using one tracking algorithm to track Arctic cyclones (ACs) throughout the manuscript. Consequently, all AC-related tables and figures have been updated. After we address this main issue, we address all other reviewer comments in order, and lastly discuss other revisions to the manuscript.

**Reviewer 2 Main Concern:**

However, there are some analysis issues I'd like to see addressed before the paper can be published. The main one being that the authors use two different cyclone tracking algorithms, while conducting their analysis. One for the detecting cyclones in reanalysis data, which is considered "truth" and another for the re-forecast data. This leads to a situation, where comparisons between the "truth" and re-forecasted cyclones are not apples to apples comparisons, but differences might arise from the use of different tracking algorithms (for example Walker, E., Mitchell, D. and Seviour, W. (2020), The numerous approaches to tracking extratropical cyclones and the challenges they present. Weather, 75: 336-341. https://doi.org/10.1002/wea.3861). I would like to see this analysis done utilizing only one tracking algorithm.

**Response:**

The reviewer brings up an important issue. Because of this issue, we decided to now use the Crawford et al. (2021) tracking algorithm for both the reanalysis dataset (i.e., ERA-Interim) and the reforecast dataset (i.e., the GEFS reforecast dataset version 2) instead of using the Sprenger et al. (2017) cyclone dataset for the reanalysis dataset and the Crawford et al. (2021) tracking algorithm for the reforecast dataset as previously done. This means that we have reconstructed the entire 2007–2017 AC climatology using the Crawford et al. (2021) tracking algorithm and have recreated all AC-related tables and figures. See L232–235 of the revised manuscript for updated text discussing our use of the Crawford et al. (2021) tracking algorithm. Note that the Crawford et al. (2021) tracking algorithm requires sea level pressure (SLP) data that is inputted into the algorithm to be regridded to a north-polar Lambert azimuthal equal-area grid. We make sure to clarify the regridding done for the ERA-Interim SLP data on L233–235 of the revised manuscript and for the ensemble SLP data on L331–333 of the revised manuscript. Also, we added “every 6 h over the Northern Hemisphere” on L330 of the revised manuscript for parallel construction with same text on L233 of the revised manuscript.

The main results of the paper concerning the AC-related tables and figures are generally consistent between Crawford and Sprenger. We now compare the original AC-related tables and figures based on Sprenger from the original manuscript, which are reproduced below in the current document, to the new tables and figures based on Crawford, which are contained in the revised manuscript. Note that we added a new figure in the revised manuscript (Fig. 2 of the revised manuscript), which will be discussed later in the current document, so most figure numbers in the revised manuscript are not the same as the figure numbers in the original manuscript. Changes to figure numbers will be indicated throughout the current document.

**Comparing Table 1 of original manuscript with Table 1 of revised manuscript**

The number of ACs during the 2007–2017 climatological period, ACs during low-skill periods and ACs during high-skill periods are lower for Crawford (Table 1 of the revised manuscript) compared to Sprenger (Table 1 of the original manuscript). Still, the main result that there is a higher number of ACs during low-skill periods compared to high-skill periods is consistent between Crawford and Sprenger. The text on L244–246 of the revised manuscript referring to the number of ACs was updated to reflect Table 1 of the revised manuscript.

|  |  |  |  |
| --- | --- | --- | --- |
|  | The 2007–2017 climatological period | Low-skill periods | High-skill periods |
| Number of days | 1067 | 328 | 281 |
| Number of ACs | 730 | 298 | 208 |

Table 1 of original manuscript. Number of days and ACs during the 2007–2017 climatological period, low-skill periods, and high-skill periods.

**Comparing Table 2 of original manuscript with Table 2 of revised manuscript:**

There are some differences in the number of ACs during the 2007–2017 climatological period, ACs during low-skill periods, and ACs during high-skill periods for which forecast skill can be evaluated at forecast lead times of 1–7 days, every 1 day, between Crawford and Sprenger, which is expected. Still, there is general consistency between Crawford and Sprenger in terms of which forecast lead times have a relatively higher or lower number of ACs among the forecast lead times.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Forecast lead time (days) | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| The 2007–2017 climatological period | 483 | 535 | 543 | 531 | 491 | 430 | 338 |
| Low-skill periods | 193 | 219 | 226 | 220 | 204 | 168 | 140 |
| High-skill periods | 134 | 149 | 154 | 142 | 138 | 132 | 99 |

Table 2 of original manuscript. Number of ACs during the 2007–2017 climatological period, ACs during low-skill periods, and ACs during high-skill periods for which forecast skill can be evaluated at forecast lead times of 1–7 days, every 1 day.

**Comparing Figs. 5 and 6 of original manuscript with Figs. 6 and 7, respectively, of revised manuscript:**

The track frequency patterns, and regions of relatively high and low track frequency, for ACs during the 2007–2017 climatological period, ACs during low-skill periods, and ACs during high-skill periods are consistent between Sprenger and Crawford, even if the track frequency values are not exactly the same (compare Fig. 5 of the original manuscript with Fig. 6 of the revised manuscript). Track frequency difference patterns are broadly consistent between Sprenger and Crawford (compare Fig. 6 of the original manuscript with Fig. 7 of the revised manuscript).

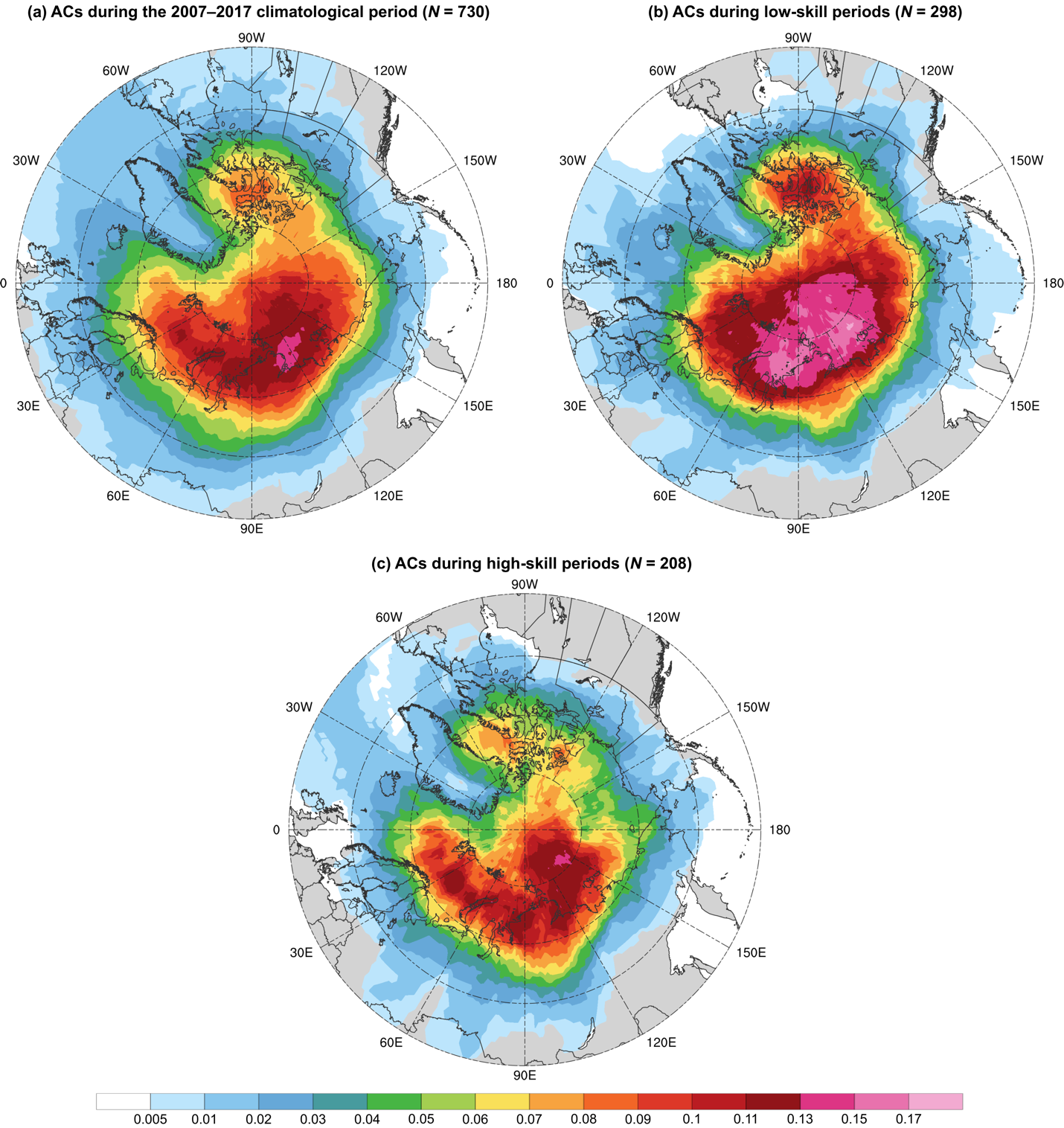


Fig. 5 of original manuscript. Track frequency of (a) ACs during the 2007–2017 climatological period, (b) ACs during low-skill periods, and (c) ACs during high-skill periods, shaded according to the number of the respective ACs for which a given grid point (using a 1° grid) is located within 500 km of the center of the respective ACs normalized by the number of days in the respective periods (given in Table 1). Units: number of ACs day−1.

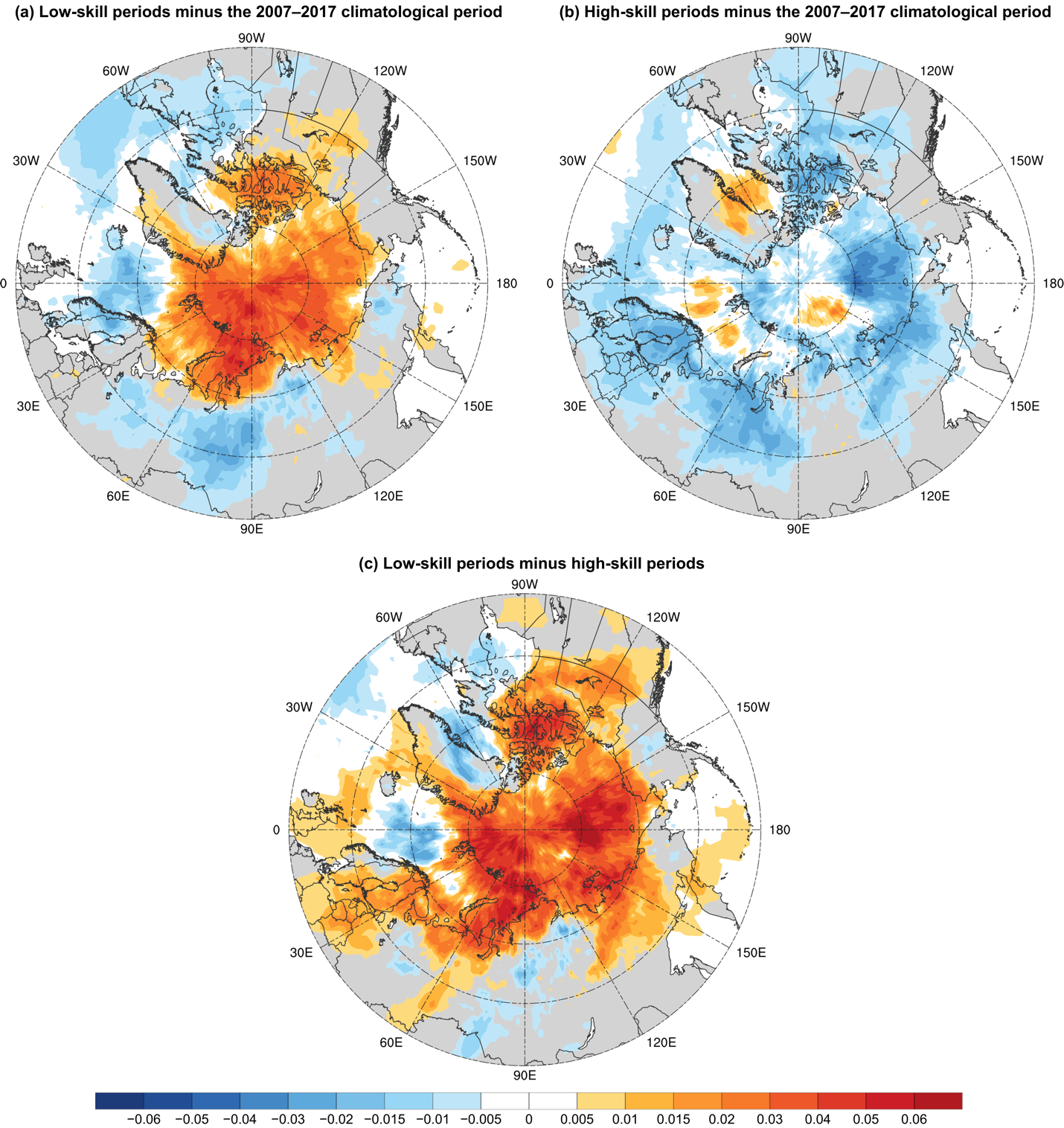


Fig. 6 of original manuscript. Difference in track frequency between (a) ACs during low-skill periods and ACs during the 2007–2017 climatological period (Fig. 5b minus Fig. 5a), (b) ACs during high-skill periods and ACs during the 2007–2017 climatological period (Fig. 5c minus Fig. 5a), and (c) ACs during low-skill periods and ACs during high-skill periods (Fig. 5b minus Fig. 5c; also Fig. 6a minus Fig. 6b). Units: number of ACs day−1.

**Comparing Figs. 8 and 9 of original manuscript with Figs. 9 and 10, respectively, of revised manuscript:**

There is consistency between Sprenger and Crawford in terms of how the distributions of quantities characterizing ACs during the 2007–2017 climatological period, ACs during low-skill periods, and ACs during high-skill periods compare among these ACs (compare Figs. 8 and 9 of the original manuscript with Figs. 9 and 10 of the revised manuscript, respectively). Statistical significance is consistent between Sprenger and Crawford for all quantities, except for area-averaged IVT when comparing between ACs during low-skill periods and ACs during the 2007–2017 climatological period (compare Fig. 9c of the original manuscript with Fig. 10c of the revised manuscript).

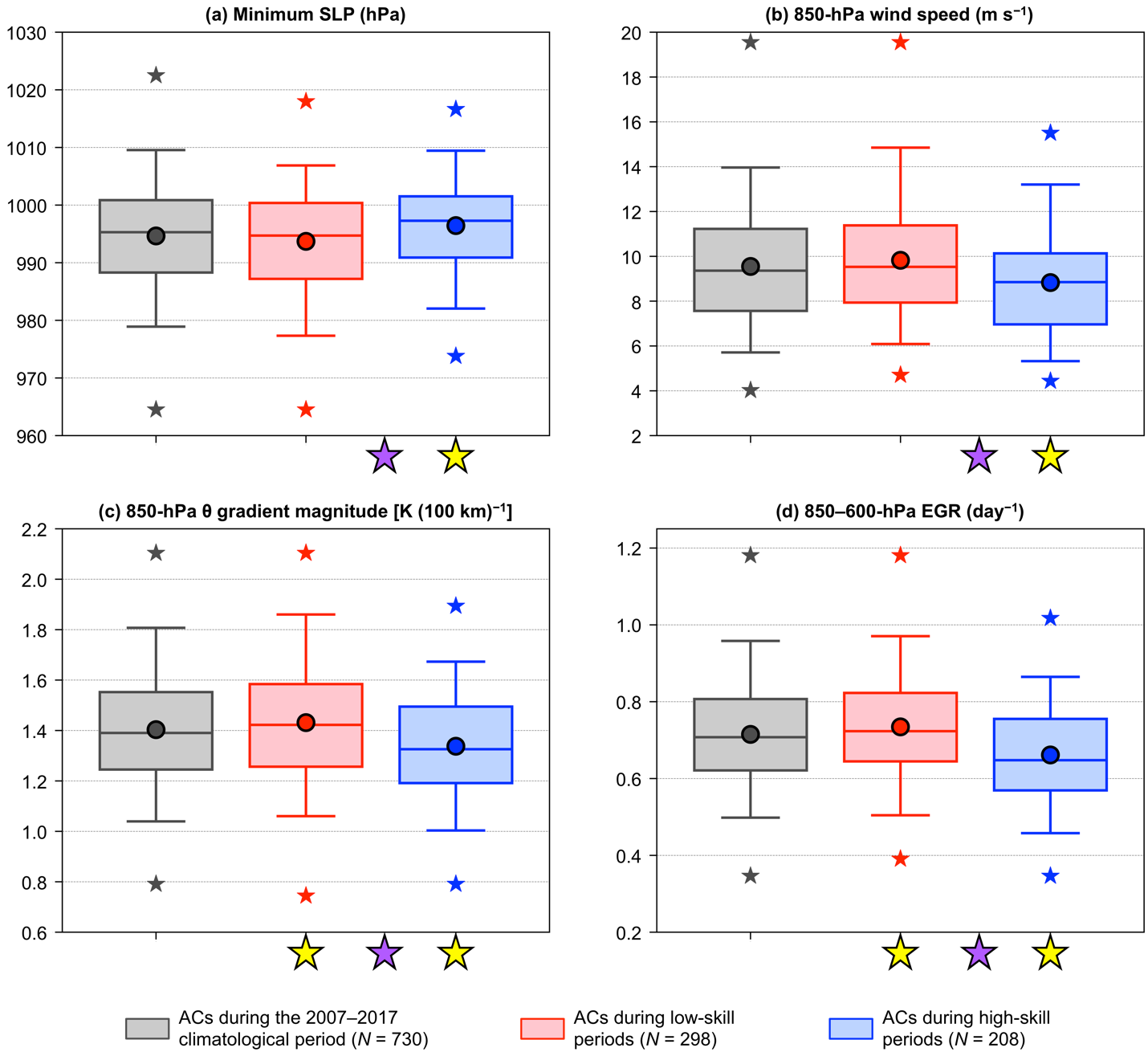


Fig. 8 of original manuscript. Distributions of the most extreme value, as defined in section 2c, of quantities characterizing ACs during the 2007–2017 climatological period (gray), ACs during low-skill periods (red), and ACs during high-skill periods (blue) when located within the Arctic during the respective periods. The selected quantities are (a) minimum SLP (hPa), (b) area-averaged 850-hPa wind speed (m s−1), (c) area-averaged 850-hPa *θ* gradient magnitude [K (100 km)−1], and (d) area-averaged 850–600-hPa EGR (day−1). The quantities in (b)–(d) are area-averaged within a 1000-km radius from the centers of the ACs. Dots indicate the mean values, boxes indicate the IQR, and whiskers extend to the 5th and 95th percentiles. Gray, red, and blue stars indicate the minimum and maximum values of the distributions. Yellow stars indicate statistical significance at the 95% confidence level of the mean values of the quantities for ACs during low-skill periods and for ACs during high-skill periods with respect to the mean values of the quantities for ACs during the 2007–2017 climatological period. Purple stars indicate statistical significance at the 95% confidence level between the mean values of the quantities for ACs during low-skill periods and the mean values of the quantities for ACs during high-skill periods.

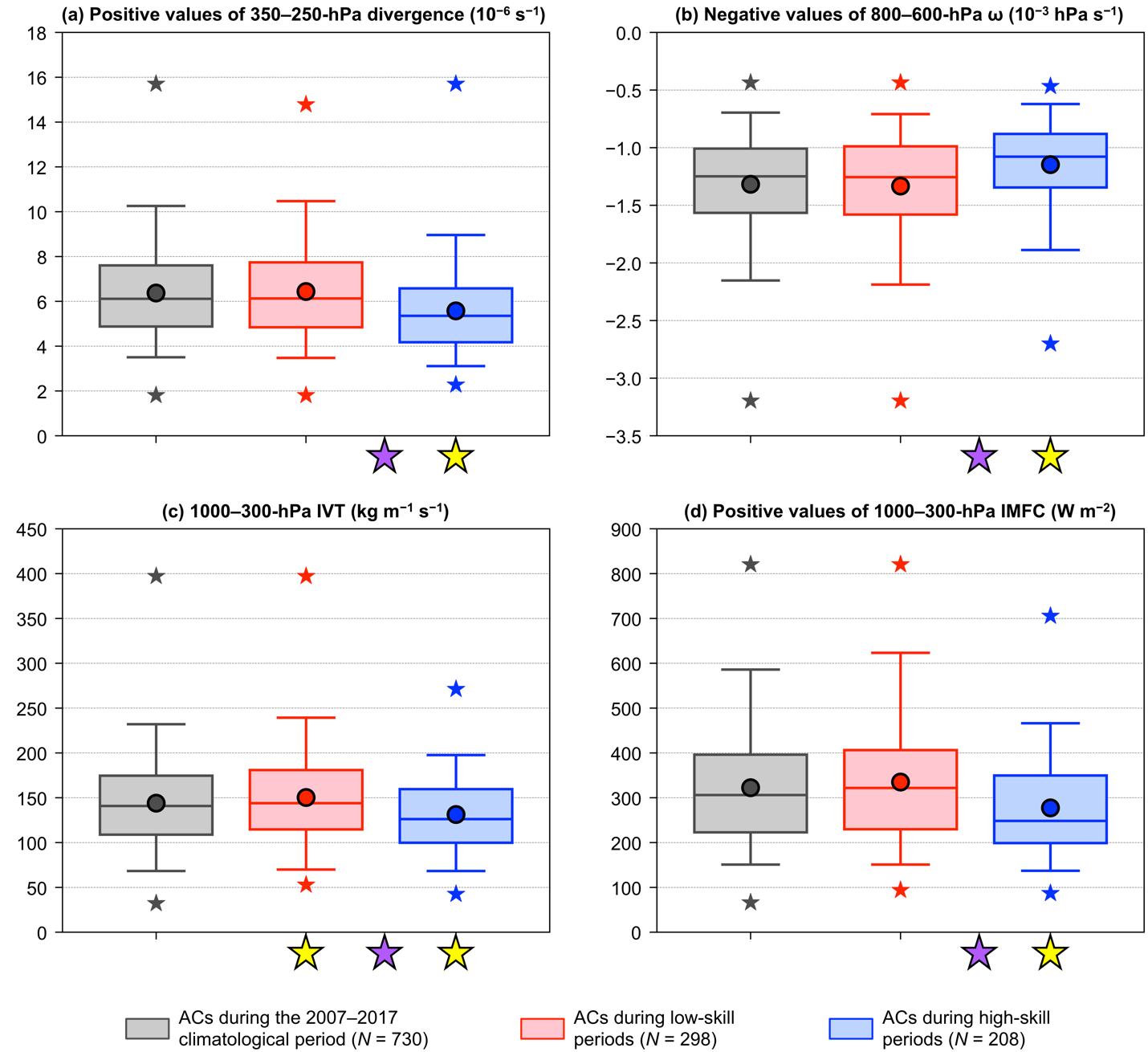


Fig. 9 of original manuscript. As in Fig. 8, but for (a) area-averaged positive values of 350–250-hPa divergence (10−6 s−1), (b) area-averaged negative values of 800–600-hPa *ω* (10−3 hPa s−1), (c) area-averaged 1000–300-hPa IVT (kg m−1 s−1), and (d) area-averaged positive values of 1000–300-hPa IMFC (W m−2). The quantities in (a)–(d) are area-averaged within a 1000-km radius from the centers of the ACs.

**Comparing Fig. 10 of original manuscript with Fig. 11 of revised manuscript:**

The mean intensity RMSE is statistically significantly higher for ACs during low-skill periods compared to ACs during high-skill periods for forecast lead times of 5–7 days for Crawford (Fig. 11a of the revised manuscript), but for forecast lead times of 1 and 6 days for Sprenger (Fig. 10a of the original manuscript). When compared to ACs during the 2007–2017 climatological period, there are more instances of statistical significance in the mean intensity RMSE of ACs during low-skill periods and ACs during high-skill periods for Crawford (Fig. 11a of the revised manuscript) than Sprenger (Fig. 10a of the original manuscript). See L674–686 of the revised manuscript for updated text regarding Fig. 11a of the revised manuscript. We also updated the sentences discussing statistical significance in the caption of Fig. 11 of the revised manuscript (see L657–666 of the revised manuscript).

In terms of position RMSE of ACs, the main result that there are no statistically significant differences in the mean values of position RMSE between ACs during low-skill periods and ACs during high-skill periods for all forecast lead times is consistent between Sprenger (Fig. 10b of the original manuscript) and Crawford (Fig. 11b of the revised manuscript). The only difference in terms of position RMSE between Sprenger and Crawford is that the mean position RMSE is statistically significantly higher for ACs during high-skill periods compared to ACs during the 2007–2017 climatological period for the 1-day forecast lead time for Crawford (Fig. 11b of the revised manuscript), but not for Sprenger (Fig. 10b of the original manuscript). See L688–690 of the revised manuscript for new text regarding Fig. 11b of the revised manuscript.

We also updated the values for the mean intensity RMSE and mean position RMSE of ACs during the 2007–2017 climatological period given on L637–641 of the revised manuscript to reflect Fig. 11 of the revised manuscript. The values given on L637–641 of the revised manuscript are generally similar between Crawford and Sprenger.

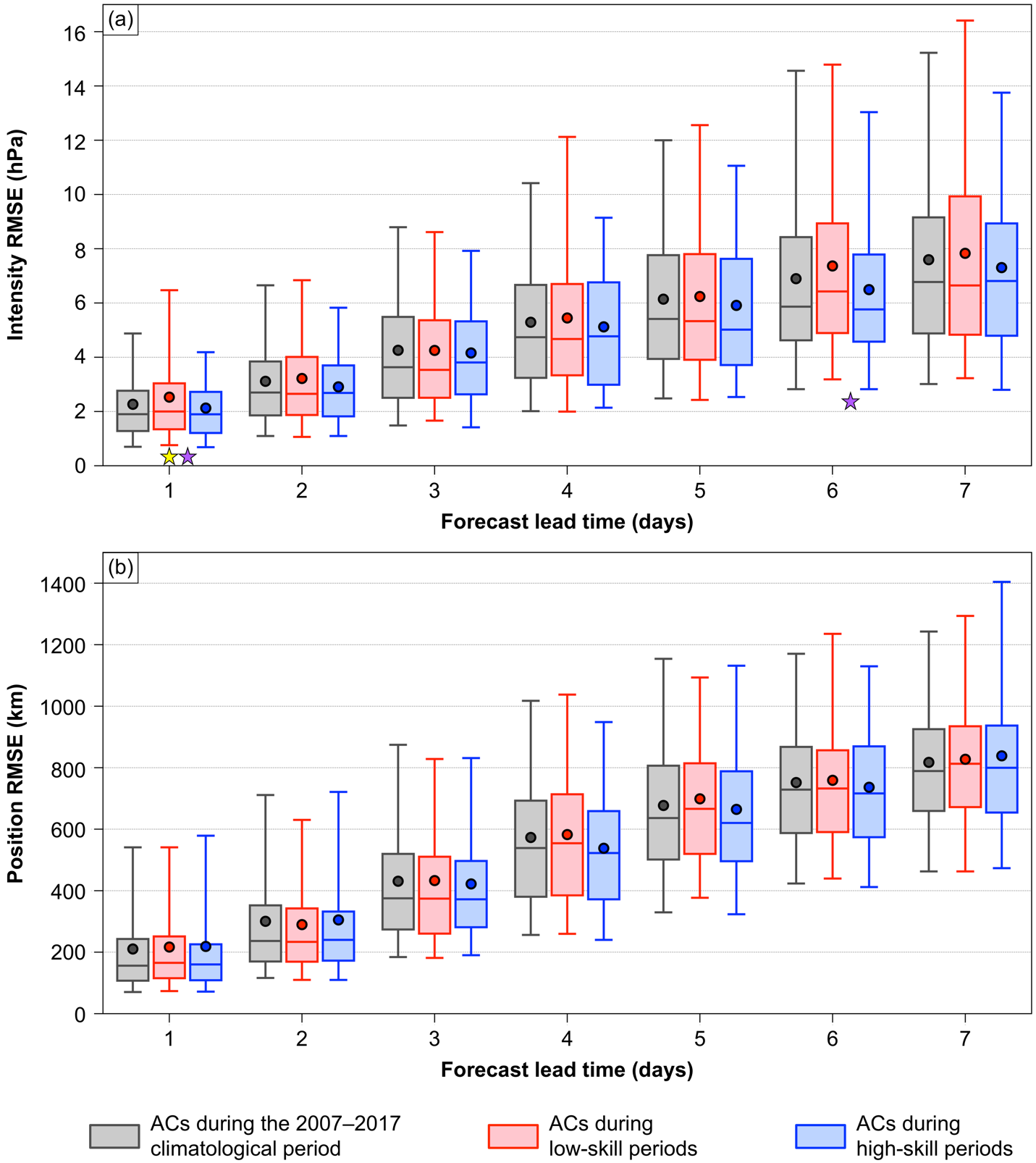


Fig. 10 of original manuscript. Distributions of (a) intensity RMSE (hPa) and (b) position RMSE (km) of ACs during the 2007–2017 climatological period (gray), ACs during low-skill periods (red), and ACs during high-skill periods (blue) for which forecast skill can be evaluated at forecast lead times of 1–7 days, every 1 day. Dots indicate the mean values, boxes indicate the IQR, and whiskers extend to the 5th and 95th percentiles. Yellow star indicates statistical significance at the 95% confidence level of the mean value of intensity RMSE for ACs during low-skill periods with respect to the mean value of intensity RMSE for ACs during the 2007–2017 climatological period for the 1-day lead time. Purple stars indicate statistical significance at the 95% confidence level between the mean value of intensity RMSE for ACs during low-skill periods and the mean value of intensity RMSE for ACs during high-skill periods for the 1-day and 6-day lead times.

**Comparing Figs. 11 and 12 of original manuscript with Figs. 12 and 13, respectively, of revised manuscript:**

These figures show distributions of quantities characterizing the four skill categories of ACs discussed in the paper. There are a greater number of instances of statistical significance for Crawford (Figs. 12 and 13 of the revised manuscript) compared to Sprenger (Figs. 11 and 12 of the original manuscript). There are some differences between Crawford and Sprenger in terms of statistically significant differences between low-skill ACs during high-skill periods and high-skill ACs during high-skill periods. For example, there is a statistically significant difference in 850–600-hPa EGR between low-skill ACs during high-skill periods and high-skill ACs during high-skill periods for Sprenger (Fig. 11d of original manuscript), but not for Crawford (Fig. 12d of the revised manuscript). See L702–709 and L732–740 of the revised manuscript for updated text regarding Fig. 12 of the revised manuscript and see L744–754 of the revised manuscript for updated text regarding Fig. 13 of the revised manuscript.

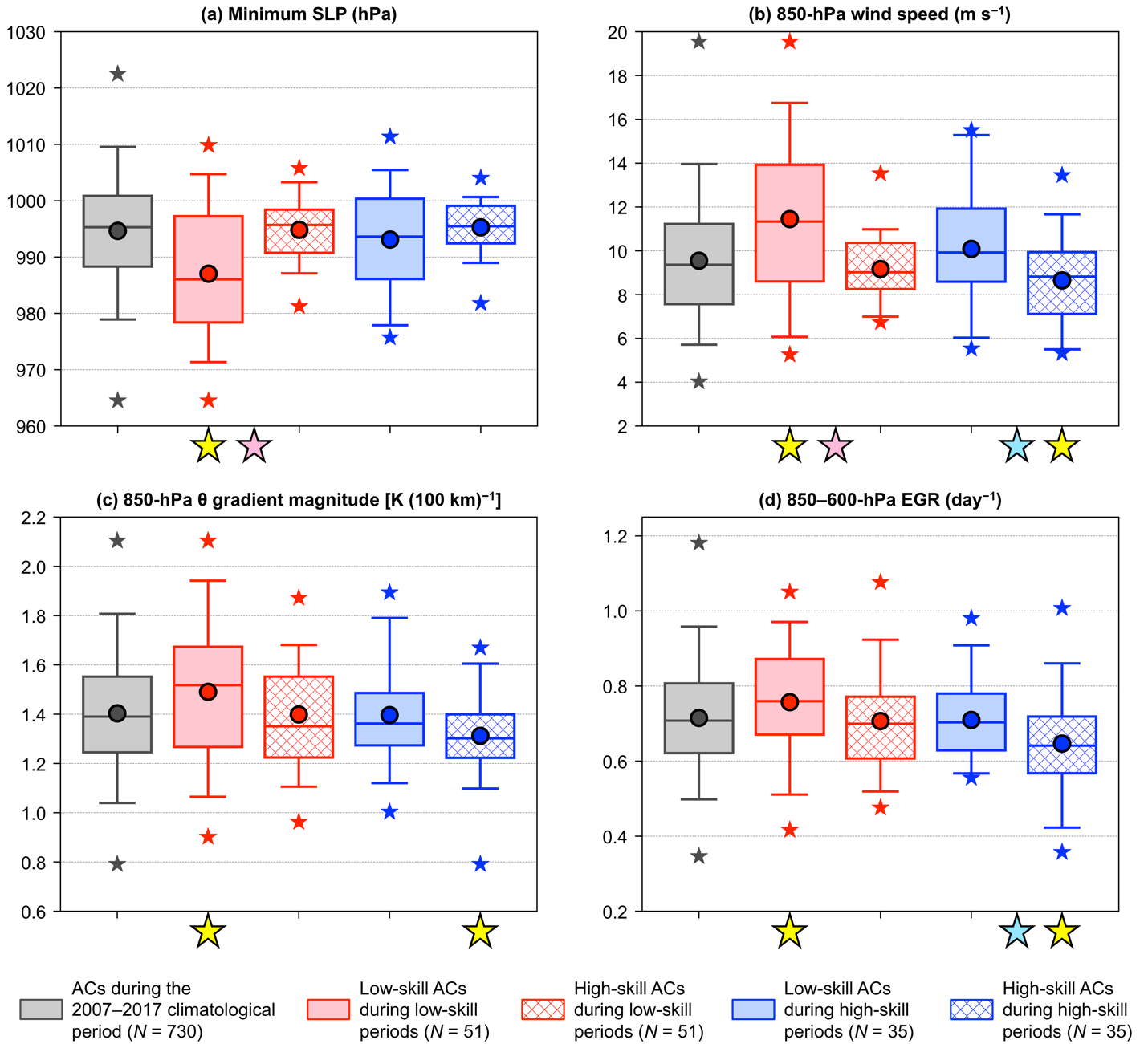


Fig. 11 of original manuscript. Distributions of the most extreme value, as defined in section 2c, of quantities characterizing ACs during the 2007–2017 climatological period (gray), low-skill ACs during low-skill periods (red solid), high-skill ACs during low-skill periods (red hatched), low-skill ACs during high-skill periods (blue solid), and high-skill ACs during high-skill periods (blue hatched) when located within the Arctic during the respective periods. The selected quantities are (a) minimum SLP (hPa), (b) area-averaged 850-hPa wind speed (m s−1), (c) area-averaged 850-hPa *θ* gradient magnitude [K (100 km)−1], and (d) area-averaged 850–600-hPa EGR (day−1). The quantities in (b)–(d) are area-averaged within a 1000-km radius from the centers of the ACs. Dots indicate the mean values, boxes indicate the IQR, and whiskers extend to the 5th and 95th percentiles. Gray, red, and blue stars indicate the minimum and maximum values of the distributions. Yellow stars indicate statistical significance at the 95% confidence level of the mean values of the quantities for low-skill ACs during low-skill periods and high-skill ACs during high-skill periods with respect to the mean values of the quantities for ACs during the 2007–2017 climatological period. Pink stars indicate statistical significance at the 95% confidence level between the mean values of the quantities for low-skill ACs during low-skill periods and the mean values of the quantities for high-skill ACs during low-skill periods. Light blue stars indicate statistical significance at the 95% confidence level between the mean values of the quantities for low-skill ACs during high-skill periods and the mean values of the quantities for high-skill ACs during high-skill periods.

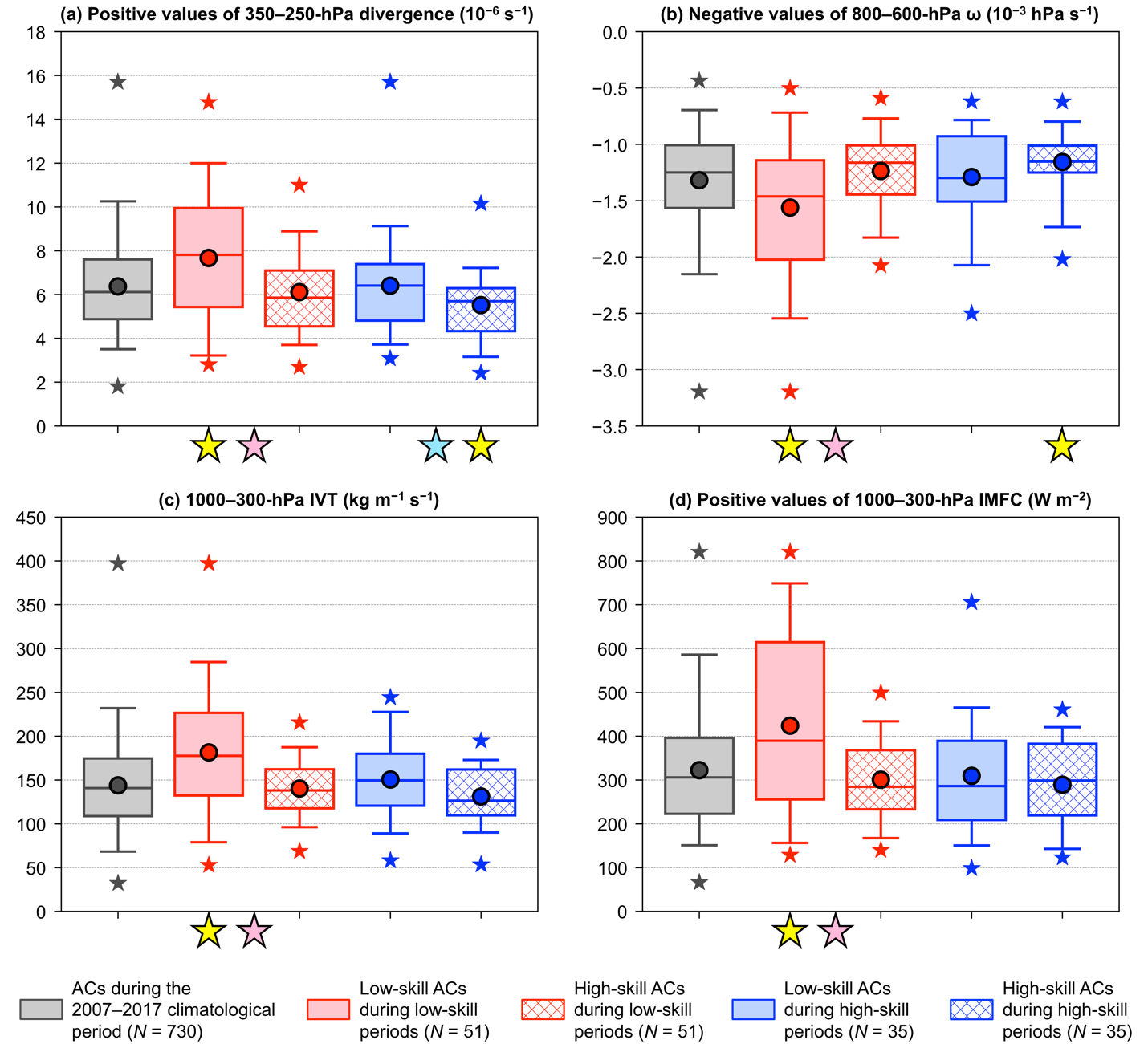


Fig. 12 of original manuscript. As in Fig. 11, but for (a) area-averaged positive values of 350–250-hPa divergence (10−6 s−1), (b) area-averaged negative values of 800–600-hPa *ω* (10−3 hPa s−1), (c) area-averaged 1000–300-hPa IVT (kg m−1 s−1), and (d) area-averaged positive values of 1000–300-hPa IMFC (W m−2). The quantities in (a)–(d) are area-averaged within a 1000-km radius from the centers of the ACs.

**Responses to Reviewer #1**

**Major Comment:**

While there is statistical significance testing performed, the authors, at times, seem to disregard the spirit of performing them in the first place. Discussions of the results frequently mix results that have statistically significant differences with those that are not statistically different (and often in the same sentence). If there is no statistical significance, the authors still discuss the differences and say something along the lines that they “tend to be [stronger/weaker/larger/etc.]” One example is on lines 629-642. It makes the results less clear and somewhat misleading to mix the statistically significant differences in with the differences that are not statistically significantly different in the discussion, and especially in the same sentences. There is a reason for performing statistical significance tests. If the authors insist on pointing out other differences that tend to occur, they should be done so separately, and it should be more strongly emphasized that there is no statistical evidence of the differences being pointed out.

**Response:**

The reviewer brings up an important concern. We now make sure to clarify which results are statistically significant and also make sure to clearly separate results that are statistically significant from results that are not statistically significant. See our responses below to your minor comments to see where changes were made.

**Minor Comments:**

1) 15-19: It seems like the purpose of this study could be framed more scientifically. I’m not sure

what we fundamentally learn by “comparing” ACs in high- vs. low-skill periods. Perhaps the purpose is to learn what processes are difficult to represent in the low-skill ACs.

**Response:**

We reframed the purpose statement in the abstract as follows:

Original purpose statement:

“The purpose of this study is to compare characteristics of the Arctic environment, and the frequency, characteristics, and forecast skill of ACs, between periods of low and high forecast skill of the synoptic-scale flow over the Arctic, hereafter referred to as low-skill and high-skill periods, respectively, during the summers of 2007–2017.”

New purpose statement (L15–17 of the revised manuscript):

“The purpose of this study is to increase understanding of processes that influence the forecast skill of the synoptic-scale flow over the Arctic and of ACs.”

We feel this new purpose statement is more scientifically framed. Because of this change, we now define low-skill periods and high-skill periods, and mention summers of 2007–2017, in the subsequent sentence on L17–21 of the revised manuscript. Also, at the end of section 1, we change “The objective of the present study is to improve understanding of Arctic environmental conditions and ACs during low-skill and high-skill periods” to the new purpose statement for consistency (see L147–148 of the revised manuscript).

2) I recommend concluding section 1 with a hypothesis. There is enough previous literature to

justify one here. For example, if Yamagami and Matsueda (2021) find that lower skill occurs

in the “Arctic Cyclone” pattern, do you expect to find the same thing in this study and why

or why not? Also, there is a discussion about how latent heating in midlatitudes provides

reason to believe latent heating could be a source of lower forecast skill. However, also given

that Capute and Torn (2021) find no differences in skill with latent heating, what do you

expect to find in this study?

**Response:**

We added hypotheses to the last paragraph of section 1 (see L150–155 of the revised manuscript).

3) 64: ACs have a summertime maximum over the central Arctic Ocean. But this is different

from saying that ACs occur most frequently during summer. In fact, by comparing Figure 1a

to Figure 1b in Serreze and Barrett (2008), it is clear there are more ACs during the winter.

**Response:**

This is a good point that was overlooked. We changed the sentence to focus on regional differences in the frequencies of ACs between winter and summer (see L63–67 of the revised manuscript).

4) 106-119: Blanchard-Wrigglesworth et al. (2022) show that the strongest cyclone on record

was well predicted up to 8 days in advance while changes in sea ice were not, and that this was

likely due to biases in the initial sea ice conditions and missing physics of wave-sea interactions

in the forecast model.

**Response:**

We added a slightly modified version of this statement and another statement concerning Blanchard–Wrigglesworth et al. (2022) to the discussion on the predictability of ACs in section 1 (see L121–125 of the revised manuscript).

5) In figures 1-4, while the y-axis lists a description and units, the x-axis only lists the units. A

similar description needs to be added to the figures for consistency.

**Reponse:**

We changed the x-axis of Figs. 3–5 of the revised manuscript (replaces Figs. 2–4 of the original manuscript) to “Forecast lead time (days)”.

6) 339-342: In figure 2a, I do not see a red dot at 0.5 days (unless it is completely overlaid with

the blue dot). So is it more appropriate to state that there is statistical significance from

days 1-5 instead of “almost all” of days 0-5.5?

**Response:**

Correct, there is no red dot at day 0.5 in Fig. 3a of the revised manuscript (replaces Fig. 2a of the original manuscript). Since there is statistical significance for day 1–5.5 for both low-skill forecasts and high-skill forecasts, we changed “for almost all of day 0–5.5” to “for day 1–5.5” on L403 of the revised manuscript.

7) 467-483: Instead of referring to the arbitrary “equatorward boundary of the Arctic,” would

it perhaps be more physical to say the AFZ? The anomalously strong shear over the AFZ

during this 500-hPa height pattern would promote stronger cyclogenesis in the AFZ region if

the low-level temperature gradient is anomalously strong.

**Response:**

We agree that it is more physical to say “AFZ” instead of “equatorward boundary of the Arctic” and have changed the terminology throughout the paragraph on L537–552 of the revised manuscript.

8) 504-507: It seems more than speculatory to point out that there is a tendency for a ridge over

the Greenland vicinity during the low-skill periods (i.e., the 552 dam contour in Figure 7a).

**Response:**

The authors agree that this point may be too speculatory. This comment pertains to a small paragraph in the original manuscript that contains discussion of the possible reason for localized regions of lower track frequency of ACs during low-skill periods compared to high-skill periods. We decided to remove the entire paragraph as the discussion is speculatory in nature and does not contribute to the main conclusions of the paper. The removed paragraph from the original manuscript is:

“There are localized regions of lower track frequency of ACs during low-skill periods over portions of Baffin Bay, western Greenland, and the Norwegian Sea (Fig. 6c), which may relate to higher 500-hPa geopotential height over these regions during low-skill periods (Fig. 7a). It is speculated that the higher 500-hPa geopotential height over the aforementioned regions during low-skill periods may indicate a tendency for increased mid-to-upper-tropospheric ridging, and concomitantly a tendency for weaker dynamical forcing for AC development, over the aforementioned regions during low-skill periods.”

9) 509-513: “Statistically significantly stronger” is mentioned four times in the sentence. This

sentence could be condensed by saying it once (i.e., ACs during low-skill periods tend to be

statistically significantly stronger (Fig. 8a), located in regions of stronger lower-tropospheric

flow (Fig. 8b), etc.).

**Response:**

In reference to Fig. 9 of the revised manuscript (replaces Fig. 8 of the revised manuscript), we kept the first two instances of “statistically significantly” in the sentence on L571–575 of the revised manuscript by stating “ACs during low-skill periods tend to be statistically significantly stronger (Fig. 9a), and tend to be located in regions characterized by statistically significantly stronger lower-tropospheric flow…”. We kept the second instance of “statistically significantly” to make it clear that the “regions of …” discussed for Figs. 9b–d of the revised manuscript are statistically significant. For the next sentence pertaining to Fig. 10 of the revised manuscript (replaces Fig. 9 of the original manuscript) on L575–578 of the revised manuscript, we state “statistically significantly” once instead of four times.

When discussing Fig. 12 of the revised manuscript (replaces Fig. 11 of the original manuscript) on L702–709 of the revised manuscript, and when discussing Fig. 13 of the revised manuscript (replaces Fig. 12 of the original manuscript) on L744–750 of the revised manuscript, we also now avoid repeatedly saying “statistically significantly” in the same sentence, and only discuss results that are statistically significant in the same sentence.

10) 669-675 (1): Very long run-on sentence beginning with “The tendency...” needs to be split

up.

**Response:**

This sentence now only discusses the statistically significant result and is thus shorter (see L737–740 of the revised manuscript). Also, given the updated AC climatology, there is now only a statistically significant difference in lower-to-midtropospheric EGR between low-skill ACs during low-skill periods and high-skill ACs during low-skill periods, which is discussed in the sentence on L737–740.

11) 669-675 (2): If there is no statistical evidence for the statement in the first half of the sentence, I would strongly recommend removing it.

**Response**

See our response to your previous comment (comment 10).

12) 685-686: Similar comment to above. The statements about stronger lower-to-midtropospheric ascent and stronger moisture transport do not have statistical significance to support it.

**Response:**

For the sentences concerning Fig. 13 in the revised manuscript (replaces Fig. 12 in the revised manuscript) on L744–750 of the revised manuscript, we now only discuss statistically significant results.

For the sentence on L750–754 of the revised manuscript that discusses comparable latent heating between low-skill ACs during high-skill periods and high-skill ACs during high-skill periods, we added “(no statistically significant difference)” on L752.

13) 738-739: If this statement supported statistically (which I think it is), then it should be stated

here instead of or in addition to stating “tendency.”

**Response:**

We added “statistically significant” before “tendency” on L801 of the revised manuscript.

14) 742: The word “dearth” was recently used; would be better to choose another word.

**Response:**

We changed “a dearth of” to “limited” on L805 of the revised manuscript.

15) 766: breakdown → decomposition

**Response:**

The corresponding sentence has been changed on L829–830 of the revised manuscript to “ACs are decomposed into four skill categories based on intensity RMSE for the 5-day lead time.” The reason for the change is discussed in our response to your next comment.

16) 766-771: The authors should carefully check to ensure these statements are only those supported with statistical significance.

**Response:**

We replaced these statements as follows:

Statements in the original manuscript:

“A breakdown of ACs into four skill categories based on intensity RMSE for the 5-day lead time reveals that low-skill ACs during low-skill periods tend to be the strongest, and tend to be located in environments most conducive to development and intensification, when compared to all other skill categories of ACs. Low-skill ACs during low-skill periods tend to be located in regions of relatively strong lower-tropospheric baroclinicity, relatively large lower-to-midtropospheric EGR, and relatively large latent heating.”

Statements in the revised manuscript (L829–834 of the revised manuscript) that replace the aforementioned statements:

“ACs are decomposed into four skill categories based on intensity RMSE for the 5-day lead time. Low-skill ACs during low-skill periods tend to be statistically significantly stronger, and tend to be located in regions of statistically significantly stronger lower-tropospheric baroclinicity, larger lower-to-midtropospheric EGR, stronger moisture transport, and larger latent heating, when compared to high-skill ACs during low-skill periods and ACs during the 2007–2017 climatological period.”

The new statements in the revised manuscript are more precisely those that are supported with statistical significance, and now also mention moisture transport. We also deleted a similar sentence from section 3 of the original manuscript, which stated: “Figures 12 and 13 suggest that low-skill ACs during low-skill periods tend to be the strongest, and tend to be located in environments most conducive to development and intensification, when compared to all other skill categories of ACs.” This sentence is similar to the first aforementioned statement from the original manuscript that was deleted.

17) It would be nice to have an even broader statement about what type of conditions forecasters

could expect to see that would indicate a higher likelihood of a poor AC forecast.

**Response:**

We added a broader statement on L837–841 of the revised manuscript.

**Responses to Reviewer #2**

**Main Issues:**

However, there are some analysis issues I'd like to see addressed before the paper can be published. The main one being that the authors use two different cyclone tracking algorithms, while conducting their analysis. One for the detecting cyclones in reanalysis data, which is considered "truth" and another for the re-forecast data. This leads to a situation, where comparisons between the "truth" and re-forecasted cyclones are not apples to apples comparisons, but differences might arise from the use of different tracking algorithms (for example Walker, E., Mitchell, D. and Seviour, W. (2020), The numerous approaches to tracking extratropical cyclones and the challenges they present. Weather, 75: 336-341. https://doi.org/10.1002/wea.3861). I would like to see this analysis done utilizing only one tracking algorithm.

**Response:**

See our response to this main issue at the beginning of this document.

In addition to the main issue, I would like to understand why the authors used ERA-Interim data to perform this analysis, even though ERA5 has been available for some time now. If there is some specific reason, it should be explicitly stated in the manuscript. Similarly, why only summer season was chosen?

**Response:**

The ERA-Interim dataset was originally used to be consistent with the use of the ERA-Interim dataset by Sprenger et al. (2017) to identify cyclones, and for ease of data availability at the time of conducting this study. Although the ERA5 dataset (Hersbach et al. 2020) could be used, the present study will focus on comparisons of area-averages of selected dynamic and thermodynamic quantities between low-skill periods and high-skill periods, which are expected to yield consistent results between ERA-Interim and ERA5 even if there are differences in the magnitude of these quantities between these datasets. We added some of this discussion on L249–254 of the revised manuscript.

We only chose the summer season to keep the period of analysis reasonably bounded for the sake of time. In addition, the summer season is examined because human activities in the Arctic like tourism (e.g., Hall and Saarinen 2010) and shipping (e.g, Eguíluz et al. 2016) are increased during summer, and strong winds, high waves, and rapid sea ice loss and movement that can be associated with strong ACs during summer (e.g., Zhang et al. 2013) may pose hazards to the human activities in the Arctic (e.g., Inoue 2021). We added some of this discussion on L175–179 of the revised manuscript.

Lastly, even though the summary section was well-written overall, it was missing discussion of any limitations of this work. This should be added.

**Response:**

We added discussion of limitations of the work and expanded discussion in the summary section on L842–868 of the revised manuscript.

**Line to line comments**

Line 45-46: change ", such that accurate…" with "which is why accurate…" to make the sentence more fluent

**Response:**

Change made on L45 of the revised manuscript.

Line 156: why this time period?

**Response:**

The reasoning for choosing summer is given in our response above to the reviewer’s main issues. The reasoning for choosing the 2007–2017 period is due to the fact that we originally wanted to also examine forecast skill over the Arctic using the ECWMF Ensemble Prediction System from TIGGE, which is only available from October 2006 onward, and due to the fact that we wanted to keep the time period of analysis reasonably bounded for the sake of time.

Line 178: why use standardized RMSs instead of "raw values"

**Response:**

We use standardized RMSs instead of “raw values” because standardized RMSs allow for a quantitative comparison of forecast skill despite differences in forecast skill that may arise from different times of the year (e.g., Torn 2017, section 2).

Reference for Torn (2017):

Torn, R. D., 2017: A comparison of the downstream predictability associated with ET and

baroclinic cyclones. *Mon. Wea. Rev.*, **145**, 4651–4672, https://doi.org/10.1175/MWR-D-

17-0083.1.

Line 243: mean radii in Valkonen et al. is calculated for the June-November season, not for the traditional June-August summer season

**Response:**

While we recognize that Valkonen et al. (2021) calculated many cyclone-related statistics for the warm season that they defined as June–November, we looked at Fig. 5a from Valkonen et al. (2021), which show distributions of average cyclone radius in the Arctic for each of the 12 months during 1979–2015, as a basis for our statement “…and Valkonen et al. (2021), who show that the median radii of ACs during June–August of 1979–2015 are approximately 1000–1100 km.”

Upon reexamining Fig. 5a from Valkonen et al. (2021), we feel that our aforementioned statement needs to be more accurately stated. Thus, on L274–276 of the revised manuscript, we changed the aforementioned statement to “…and Valkonen et al. (2021), who show in their Fig. 5a that the interquartile ranges of average cyclone radius in the Arctic for June–August of 1979–2015 for three different reanalysis datasets encompass a value of 1000 km.”

Line 267-276: "There is one candidate verification time if either…." These sentences are confusing, please rephrase

**Response:**

To make these sentences clearer, we added a schematic diagram (Fig. 2 of the revised manuscript) for the possible situations involving the identification of candidate verification times. For each situation described on L298–311 of the revised manuscript, we reference the corresponding situation shown in Fig. 2 of the revised manuscript. We also updated the text corresponding to Situation D on L306–307 of the revised manuscript from “if one of the surrounding 0000 UTC times does not occur during the period of interest” to “if the later surrounding 0000 UTC time does not occur during the period of interest” since the earlier surrounding 0000 UTC time will always occur during the period of interest.

Line 277-287: To my understanding the authors used different cyclone tracking algorithms to detect the cyclones in the ERA-Interim analysis and in the ensemble forecasts. This is problematic as two different tracking algorithms will provide different results even when ran on same data. If this is not the case and I'm misunderstanding something, please explain. Otherwise, the analysis should be done with one tracking algorithm. More on this in the general comments.

**Response:**

See our response to this main issue at the beginning of this document.

Line 277-287: How close do the cyclones need to be to each other (spatially and temporally) to be considered "matched". Please provide these numbers in the text.

**Response:**

We added discussion on these numbers on L340–346 of the revised manuscript.

Line 301: what is great circle distance?

**Response:**

The great circle distance is the shortest distance between two points on a sphere (e.g.,

<https://mathworld.wolfram.com/GreatCircle.html>). The great circle distance can be used to determine track forecast error of features (e.g., used by the National Hurricane Center to determine track forecast error of tropical cyclones: <https://www.nhc.noaa.gov/verification/verify2.shtml>).

Line 340-341: whenever you are listing high and low skill results together, please use brackets to make the sentence clearer. For example on line 340 "area-averaged RMSE is statistically significantly high and low relative to climatology for low-skill forecasts and high-skill forecasts, respectively,…" replace with "area-averaged RMSE is statistically significantly high (low) relative to climatology for low-skill (high-skill) forecast, respectively,…"

**Response:**

We decided to avoid parenthetical construction. Our rationale is based on guidance from a Monthly Weather Review editorial by Schultz et al. (2022), in which they recommend avoiding parenthetical construction in submissions to Monthly Weather Review (page 2825).

Reference for Schultz et al. (2022):

Schultz, D. M., and Coauthors, 2022: How to be a more effective author. *Mon. Wea. Rev.*, **150**, 2819–2828, https://doi.org/10.1175/MWR-D-22-0277.1.

Line 342: same, replace "…increases and decreases throughout most of day 0-5.5 for low-skill forecasts and high-skill forecasts, …" with "increases (decreases) throughout most of day 0-5.5 for low-skill (high-skill) forecasts,"

**Response:**

Same response as before regarding parenthetical construction.

Line 344-348: do similar replacements for all the high-low skill comparison pairs

**Response:**

Same response as before regarding parenthetical construction.

Line 403-414: very interesting results and good summary of the findings and what they might mean.

**Response:**

We thank the reviewer for their feedback.

Line 573: remove "…for which forecast skill can be evaluated at forecast lead times of 1-7 days, every 1 day." This is repetition of the sentence above.

**Response:**

We replaced the quoted portion of the sentence with “for these forecast lead times” on L634 of the revised manuscript to shorten the sentence while still clarifying that we are referring to the number of ACs for these forecast lead times.

Figures:  
Figure 3: In the caption change all "distributions" references to "timeseries". Distributions mean histograms or pdfs and these figures are timeseries. Also shorten the caption text to only include the information needed to understand the figure, more like in figure 4.

**Response:**

We changed all “distributions” references to “timeseries” in the caption of Fig. 4 of the revised manuscript (replaces Fig. 3 of the original manuscript).

We shortened the caption of Fig. 4 of the revised manuscript by moving all text in the last two sentences beginning with “Days 0–5.5 for low-skill forecasts and high-skill forecasts..” to the body text (see L425–431 of the revised manuscript) as we still want to keep this text in the paper.

Figure 5: Could you rephrase how the track frequencies are calculated? Are tracks calculated for each of the 1 degree grids or over 500km area? Or do you mean that each cyclone that is detected, is calculated to exist at each grid point 500 km from the central point?

**Response:**

Track frequency is calculated at each grid point. For an AC during a period of interest (e.g., low-skill periods), as long as the center of the AC is located within 500 km of a given grid point at any time during the period of interest, the AC is counted at that grid point. It is correspondingly true that for a given AC during a period of interest, the AC is calculated to exist at each grid point 500 km from the central point of the AC. We rephrased the caption of Fig. 6 of the revised manuscript (replaces Fig. 5 of the original manuscript) on L512–517 of the revised manuscript.

Figure 6: This figure seems to be a bit unnecessary. You might want to consider moving it to appendices or getting rid of it completely. As your results comparing the track densities are so clear, which is great!, this figure appears just to be a repetition of the information on figure 5.

**Response:**

We thank the reviewer for their feedback. We still think that Fig. 7 of the revised manuscript (replaces Fig. 6 of the original manuscript) still helps the reader to more easily visualize the differences in track frequency that are discussed, so we decided to keep this figure in the body text.

Figure 7: I would change the order at which the authors explain the black contours vs shading. At least for me, I was first looking at the shading and having that explained first would be helpful to make the figure more easily understandable.

**Response:**

We now explain the shading first in the caption of Fig. 8 of the revised manuscript (replaces Fig. 7 of the original manuscript). We also simplified the sentence regarding black contours by now just stating “Black contours show” at the beginning of the sentence instead of repeatedly stating black contours for each field shown in Figs. 8a–c of the revised manuscript. See L563–568 of the revised manuscript.

**Responses to Reviewer #3**

**Main Comments:**

Lines 170-176: Does the calculation of RMSE is correct in Eq (1)? The authors wrote that the RMSE was calculated at each grid point, and then the ensemble mean RMSE at each grid point was calculated, at last, the values were averaged over the Arctic (>70N) next. Is the area-averaged RMSE equal to the RMSE of the ensemble mean forecast (the ensemble mean of predicted Z500 was calculated first, and then the RMSE was calculated at each grid point, and at last the RMSE was averaged over the Arctic)? The reviewer guesses that the definition of area-averaged RMSE might contain forecast error of each member and ensemble spread. Thus, I am confused about the meaning of the area-averaged RMSE. Could you please add some explanation about the meaning of area-averaged RMSE?

**Response:**

To clarify, the RMSE is first calculated at each grid point over the Arctic according to Eq (1) on L186 of the revised manuscript, and then the RMSE values at all grid points over the Arctic are area-averaged to obtain an area-averaged RMSE value. There is no step in which “the ensemble mean RMSE at each grid point was calculated” after calculating RMSE according to Eq (1) at each grid point. The area-averaged RMSE we calculate is thus not the same as the RMSE of the ensemble mean forecast in which the ensemble mean of predicted Z500 is calculated first, and then the RMSE is calculated at each grid point, and at last the RMSE is averaged over the Arctic.

The area-averaged RMSE we calculate contains the forecast error of individual ensemble members at individual grid points [as calculated by Eq (1)] that is area-averaged over the Arctic. The method we use is not the same as the RMSE of the ensemble mean forecast, which is more comparable to the ensemble spread. The method we use for calculating RMSE in Eq (1) is similar to the method used, for example, by Fang et al. (2011) (their equation 2 in section 2b).

We added the following information to L191–193 of the revised manuscript: “Area-averaged RMSE contains the forecast error of individual ensemble members at individual grid points that is area-averaged over the Arctic.”

Reference:

Fang, X., Y.-H. Kuo, and A. Wang, 2011: The impact of Taiwan topography on the predictability of Typhoon Morakot’s record-breaking rainfall: A high-resolution ensemble simulation. *Wea. Forecasting*, **26**, 613–633, <https://doi.org/10.1175/WAF-D-10-05020.1>.

Line 178: The reviewer guesses that the forecast skill of GEFS reforecast has an increasing trend, which is due to the increase in the number of observations. Do the standardized anomalies of area-averaged RMSE (𝜎𝑅𝑀𝑆𝐸) have a trend in the period of 1985-2017? If the 𝜎𝑅𝑀𝑆𝐸 has a trend, does the trend has no influence on the results? (For example, most of the low-skill (high-skill) period tends to appear in the first (latter) half of the whole period.)

**Response:**

We plotted a timeseries of σRMSE averaged for JJA of each year during 1985–2017 (see Fig. 1 given below in the current document). We find a general negative trend during 1985–2017, with a negative slope of a least squares regression line calculated for the 1985–2017 period (slope: −0.014 σ/year). The negative trend suggests that there is some increase of forecast skill over the Arctic in the GEFS reforecast dataset version 2 over the 1985–2017 period, though there is considerable interannual variability of σRMSE during 1985–2017.

We also plotted a time series of the number of low-skill forecasts and high-skill forecasts identified for JJA of each year during 2007–2017 (see Fig. 2 given below in the current document). Fig 2 in the current document suggests that there is somewhat of a positive and negative trend in the number of low-skill forecasts and high-skill forecasts, respectively, during 2007–2017. However, there is considerable interannual variability in the number of low-skill forecasts and high-skill forecasts, with the trends in the number of low-skill forecasts and high-skill forecasts likely influenced by a couple years with a relatively high number of these forecasts. The positive trend for low-skill forecasts appears to be influenced by a relatively large number of low-skill forecasts during 2016 and 2017 and the negative trend for high-skill forecasts appears to be influenced by a relatively large number of high-skill forecasts during 2007. Overall, the negative trend of σRMSE during 1985–2017 does not appear to influence when most low-skill and high-skill forecasts occur during 2007–2017.

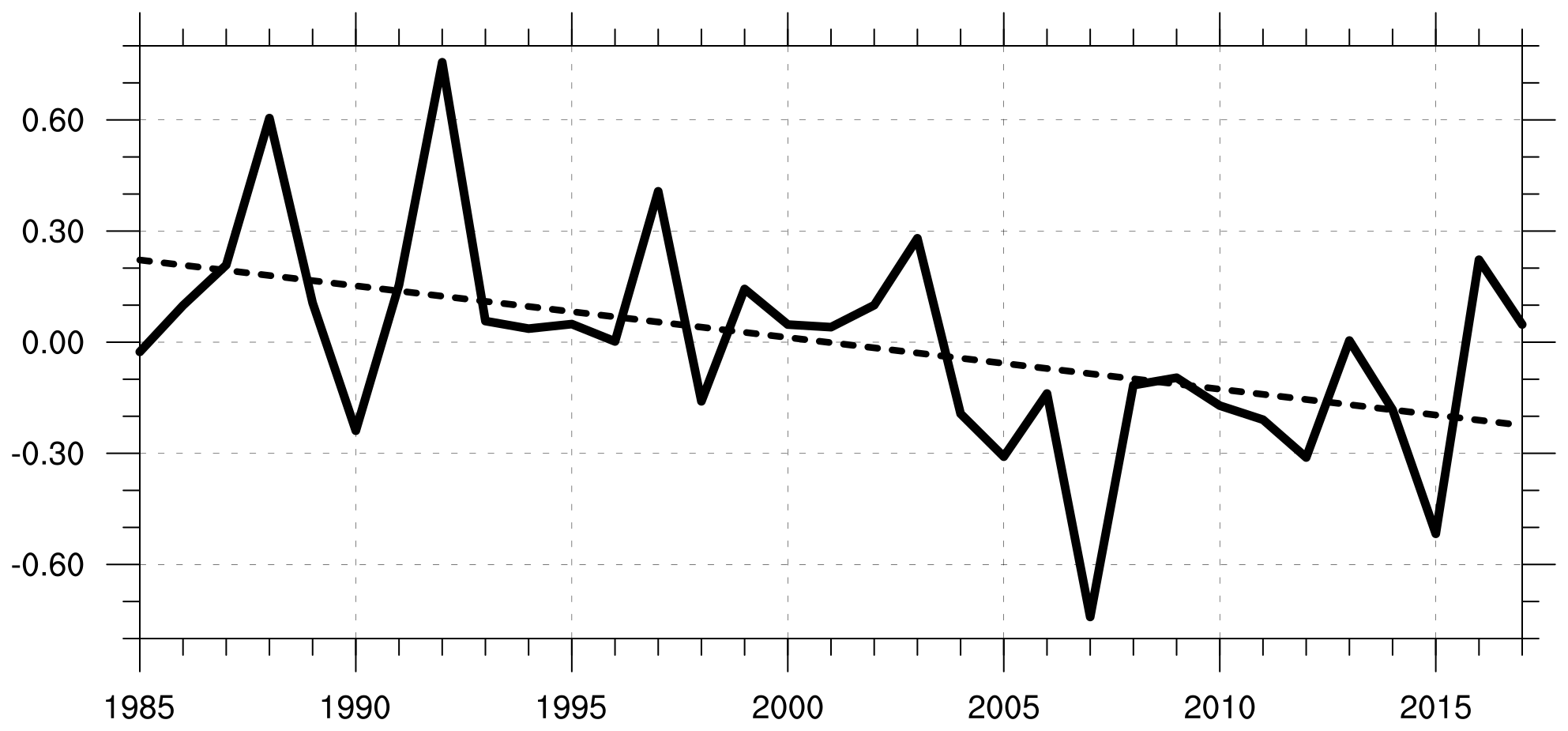


Fig. 1. Timeseries of σRMSE of 500-hPa geopotential height over the Arcticaveraged for JJA of each year during 1985–2017 (σ, solid line) and least squares regression line calculated for the 1985–2017 period (dashed line).

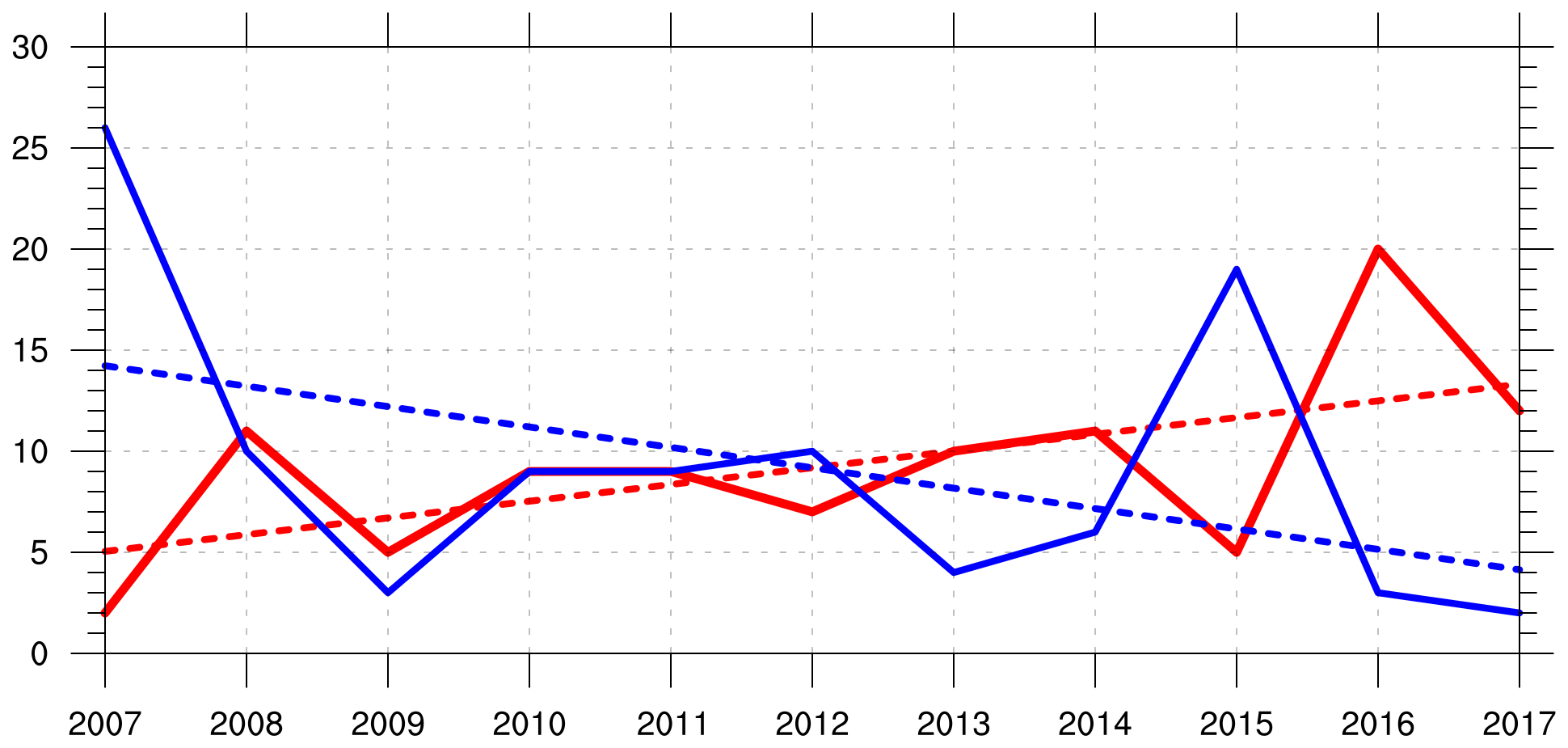


Fig. 2. Timeseries of the number of low-skill forecasts (solid red line) and high-skill forecasts (solid blue line) for JJA of each year during 2007–2017, and least squares regression line calculated for the 2007–2017 period for low-skill forecasts (dashed red line) and high-skill forecasts (dashed blue line).

Line 185-186, Fig. 1b: Although the authors wrote that the 𝜎𝑅𝑀𝑆𝐸 was calculated from the 5-day forecast, figure 1 showed 𝜎𝑅𝑀𝑆𝐸 at lead times from 0 to 5.5 days, every 0.5 days. Does the 𝜎𝑅𝑀𝑆𝐸 was calculated based on each forecast lead time?

**Response:**

Correct, we did calculate area-averaged RMSE and σRMSE for all forecast lead times of 0 to 5.5 days, every 0.5 days, in addition to just the forecast lead time of 5 days, to show the evolution of forecast skill as a function of forecast lead time. To make this clear, we added this information to L200–203 of the revised manuscript.

Line 196-192: The definition of high- and low-skill periods are based on the 10th percentile of the σ\_RMSE from daily forecasts. However, the authors removed the overlapping period. Is there a threshold for the number of successive days for the high- and low-skill period? the reviewer guesses that the removal of overlapping days corresponds to the removal of forecast bust-like events (high- (low-) skill forecast initialized on a day and low- (high-) skill forecast initialized on the next day). Could you please explain in more detail about removing overlapping days?

**Response:**

There is no threshold for the number of successive days for low-skill and high-skill periods. As long as there is any overlap between a low-skill period and a high-skill period, all days that overlap are considered overlapping days that are assigned to the period that occurs earlier. As an example of an instance in which there are overlapping days, a high-skill period was identified to occur during 8–13 June 2010 and a low-skill period was identified to occur during 12–17 June 2010, making 12–13 June 2010 overlapping days. Since the high-skill period occurs earlier than the low-skill period, 12–13 June 2010 are assigned to the high-skill period and removed from the low-skill period. We added some of this discussion on L213–218 of the revised manuscript.

Although this methodology removes a portion of a low-skill period or a high-skill period that overlap, day 5 of each period (for which σRMSE is in the top or bottom 10%) is always at least least retained for each period. Hypothetically, if the high-skill period occurred during 8–13 June 2010 (forecast initialized on 8 June) and the low-skill period occurred during 9–14 June 2010 (forecast initialized on 9 June), 9–13 June would be considered overlapping days that would be assigned to the high-skill period and removed from the low-skill period. Day 5 of each period (13 June for the high-skill period and 14 June for the low-skill period) would still be retained.

Line 228-229: Dose the “positive” ("negative") means the average of only positive (negative) values area? (Does it correspond to the area average over the Arctic and within a 1000km radius?)

**Response:**

Correct, the “positive” (“negative”) means the average of only positive (negative) values area. This corresponds to the area average over the Arctic and within the 1000-km radius of the ACs.

Line 212-213: Are there ACs which existed in the Arctic through the two periods (e.g., climatological period to high-skill period)? How the authors classified such ACs?

**Response:**

A given AC can be identified for more than one of the periods (e.g., low-skill periods and high-skill periods) as long as the AC occurs at any time within the Arctic during these periods. Furthermore, since low-skill periods and high-skill periods are subsets of the 2007–2017 climatological period, ACs during low-skill periods and ACs during high-skill periods are subsets of ACs during the 2007–2017 climatological period. This text was added to L240–244 of the revised manuscript.

Also, on L238 of the revised manuscript, we added “within the Arctic” after “All ACs that occur at any time” because “within the Arctic” was accidentally left out of the original manuscript.

Line 269-276: The reviewer feels that the selection of the verification time was confusing. If possible, a schematic figure can help the understanding of the selection of verification time.

**Response:**

As discussed in our response to a similar comment from Reviewer 2 on L621–628 of the current document, we added a schematic diagram (Fig. 2 of the revised manuscript) for the possible situations involving the identification of candidate verification times.

Line 277: Is the influence of the difference in cyclone tracking methods between ERA-interim (Sprenger et al., 2017) and GEFS (Crawford et al., 2021) small?

**Response:**

See our response to the main issue from Reviewer 2 at the beginning of the current document, where we discuss the influence of the choice of tracking algorithm on the results of the manuscript.

Line 291: The reviewer thinks that the matching criteria (≥5 ensemble members) ignore the cases of forecast deterioration when the AC existence did not predict correctly at the verification time. Were the cases ignored or classified into the low-skill AC?

**Response:**

The matching criteria (≥ 5 ensemble members) ignores cases in which < 5 ensemble members have a matching forecast AC occurring at the verification time. This was done because we want to ensure we have enough ensemble members to properly calculate intensity RMSE and position RMSE. We specifically focused on examining ACs characterized by low and high forecast skill of intensity. We did not classify the ignored cases as ACs characterized by low forecast skill of intensity since the forecast skill of intensity could not be calculated for the ignored cases.

ACs for which there are < 5 ensemble members with a matching forecast AC occurring at the verification time could be ACs characterized by low forecast skill of existence. There is a higher number ACs characterized by low forecast skill of existence during low-skill periods compared to high-skill periods for forecast lead times of 1–7 days (not shown), suggesting that low-skill periods may be associated with lower forecast skill of the existence of ACs compared to high-skill periods. We added some of this discussion in a new paragraph in section 4 that discusses limitations of the present study (see L853–860 of the revised manuscript).

Line 342-344: Does the negative value of the 𝜎𝑅𝑀𝑆𝐸 for the climatological period indicated the trend of forecast skill? In addition, why is the variability of 𝜎𝑅𝑀𝑆𝐸 (shading in Fig. 1b) largest at lead times of 0-0.5 days and does it decrease with lead time?

**Response**

We think that the negative values of σRMSE for the 2007–2017 climatological period in Fig. 3b of the revised manuscript (replaces Fig. 2b of the original manuscript) indicate that the forecast skill over the Arctic is higher during 2007–2017 relative to the 1985–2017 climatology. Figure 1 of the current document, as discussed in our response to your earlier comment, does indicate a general negative trend of σRMSE during 1985–2017, which indicates generally higher forecast skill over the Arctic during 2007–2017 compared to earlier years.

Variability of σRMSE [shading in Fig. 3b of the revised manuscript (replaces Fig. 2b of the original manuscript)] being largest at lead times of 0–0.5 days and decreasing with lead time for low-skill forecasts and high-skill forecasts is likely due to the fact that the low-skill forecasts and high-skill forecasts are determine based on σRMSE at day 5. As discussed in section 2a, forecasts associated with the top and bottom 10% of σRMSE at day 5 are referred to as low-skill forecasts and high-skill forecasts, respectively. Thus at day 5, we would expect relatively small variability of σRMSE. At earlier forecast lead times during low-skill forecasts and high-skill forecasts, σRMSE is not required to be in the top and bottom 10%, respectively, such that there is greater variability of σRMSE at earlier forecast lead times. For example, there could be one low-skill forecast with relatively low forecast skill at day 0 (relatively high value of σRMSE at day 0) but a second low-skill forecast with relatively high forecast skill at day 0 (relatively low value of σRMSE at day 0). Since both forecasts are low-skill forecasts, forecast skill will degrade over time such that σRMSE in both forecasts is in the top 10% at day 5.

Lines 364-366: Does the “anomalous” indicate the 25-75th percentile values separate from the climatological values, or indicate the statistically significant difference between high- (low-) skill forecasts and climatological period?

**Response:**

“Anomalous” indicates the statistically significant differences between high- (low-) skill forecasts and the 2007–2017 climatological period.

Line 414-419: As the authors mentioned, the vigorous baroclinic situation and latent heating would relate to the lower forecast skill over the Arctic. Although the vigorous baroclinic situation and latent heating persist throughout 0-5.5 days, the peak of the situation appeared around 3.5-4.5 days, especially wind speeds, divergence, IVT, and IMFC. Thus, the reviewer is interested in which is relatively more important. Additional discussion about that point would be helpful.

**Response:**

From the research we have done, we cannot know exactly which forecast errors (e.g., forecasts errors related to baroclinic processes and forecast errors related to latent heating) and at which times contribute most to lower forecast skill over the Arctic. We anticipate that the times of most vigorous baroclinic processes and latent heating may be associated with relatively large forecast errors related to baroclinic processes and latent heating, respectively, that contribute to low forecast skill of the synoptic-scale flow over the Arctic. We added the following discussion on L482–487 of the revised manuscript at the end of section 3a:

Relatively large lower-to-midtropospheric EGR for low-skill periods during days 1.5–3.5 (Fig. 4d) and relatively large latent heating for low-skill periods during days 2.5–4.5 (Fig. 5d) may suggest that there are relatively large forecast errors related to baroclinic processes and latent heating during these respective times that contribute to low forecast skill of the synoptic-scale flow over the Arctic during low-skill periods.

Lines 504-507: Does the weaker dynamical forcing for AC development suggest low-skill forecasts due to other processes, such as blocking?

**Response:**

The sentence in the original manuscript that this comment pertains to is:

“It is speculated that the higher 500-hPa geopotential height over the aforementioned regions during low-skill periods may indicate a tendency for increased mid-to-upper-tropospheric ridging, and concomitantly a tendency for weaker dynamical forcing for AC development, over the aforementioned regions during low-skill periods.”

We decided that based on our response to Reviewer 1’s comment on L358–371 of the current document concerning the above sentence that we would remove the paragraph containing the above sentence.

Line 520-522: Because these variables were calculated as the average within a 1000 km radius, some part of the difference in variables was induced by the ACs (e.g., stronger ACs lead to stronger wind speeds). Thus, do the differences between high- and low-skill forecasts just represent the differences associated with the difference in ACs? How did the authors separate the environment and ACs?

**Response:**

While some parts of the differences in quantities between ACs during low-skill periods and ACs during high-skill periods are induced by the ACs themselves (such as stronger ACs leading to stronger 850-hPa wind speed), there is no way to disentangle what aspects of the differences in quantities are due to the environment and due to the ACs themselves based on our methodology described in section 2c. Our methodology does not separate the environment and ACs, such that differences in quantities between ACs during low-skill periods and ACs during high-skill periods may represent differences associated with both the environment and the ACs themselves. An AC can, for example, move into an environment characterized by strong lower-tropospheric baroclinicity and concomitantly impact the magnitude of lower-tropospheric baroclinicity. We feel that the area-averages of quantities within a 1000-km radius of ACs, as discussed in section 2c, still provide a measure to compare how baroclinic processes and latent heating may influence ACs between low-skill periods and high-skill periods.

Lines 621-628: It is interesting that both high- and low-skill periods contain both low-and high-skill ACs. The reviewer agrees that ACs are one of the reasons for the low-skill forecast over the Arctic and it is not necessarily in all low-skill forecasts. Thus, the reviewer concerns with the treatment of the cases when the AC exists in ≤4 ensemble members at the verification time and its impact on these results.

**Response:**

As discussed in our response to your previous comment on ACs with < 5 ensemble members at the verification time, we added some discussion on L853–860 of the revised manuscript.

Lines 665-: The difference in the characteristics of ACs between high- and low-skill AC in low-skill period appeared in various fields (e.g., minimum SLP, wind speed, divergence, etc.), while these differences in high-skill period appeared only in wind speed, EGR, and divergence. The reviewer guesses that these differences might depend on the stage of AC (developing, mature, and decaying stages). If the authors have some speculations about the relationship between the AC stages and differences in skill and characteristics of ACs, some additional discussion would be helpful.

**Response:**

Since we did not examine the relationship between the AC stages and differences in skill and characteristics of ACs, we added some speculation in section 4 about this relationship on L842–848 of the revised manuscript. We speculate that ACs may be associated with more vigorous baroclinic processes and latent heating during the developing and mature stages compared to the decaying stage. We speculate that the forecast skill of intensity of ACs may be lower during the developing and mature stages compared to the decaying stage due to the potential for greater forecast errors related to more vigorous baroclinic processes and latent heating during the developing and mature stages compared to the decaying stage.

Lines 690-695: How about the difference in the geophysical location of ACs in high-skill periods? The cyclone frequency is higher in high-skill periods than in low-skill periods over the North Atlantic (Fig. 6). the reviewer guesses that the differences in the characteristics of ACs would be small despite the difference in forecast skills if the predictability for the ACs over North Atlantic is essentially higher than that for the ACs over other areas.

**Response:**

Based on the research done, it is difficult to answer this question beyond speculation. Geographical location of ACs may influence the characteristics of ACs and the forecast skill of ACs. For example, if forecast skill is higher over the North Atlantic compared to other regions, ACs originating over the North Atlantic may potentially be characterized by higher forecast skill compared to ACs originating over other regions. There concomitantly may be differences in preferred characteristics between ACs originating over the North Atlantic and ACs originating over other regions. We added the following discussion to L848–853 of the revised manuscript in section 4:

Another limitation of the present study is that the influence of the geographical location of ACs on the forecast skill and characteristics of ACs is not considered. It is speculated, for instance, that ACs originating from lower-latitude regions with relatively large coverage of observational data may tend to be associated with higher forecast skill compared to ACs originating over the Arctic, where there is relatively sparse coverage of conventional data (e.g., Lawrence et al. 2019).

Line 696-698: Capute and Torn (2021) compared the forecast skills between the summertime ACs and wintertime Atlantic cyclones. Does the difference in the results come from the comparison between high- and low-skill ACs in low-skill periods (both of these are summertime ACs)?

**Response:**

While Capute and Torn (2021) do compare the forecast skills between summertime ACs and wintertime Atlantic cyclones, they also compare ACs characterized by lower predictability in terms of intensity and ACs characterized by higher predictability in terms of intensity during summer. The results of Capute and Torn (2021) discussed on L757–759 of the revised manuscript concern these ACs characterized by lower predictability in terms of intensity and ACs characterized by higher predictability in terms of intensity during summer. We speculate that differences in methodology used to determine and examine low-skill ACs during low-skill periods and high-skill ACs during low-skill periods in the present study, and used to determine and examine ACs characterized by lower predictability in terms of intensity and ACs characterized by higher predictability in terms of intensity in Capute and Torn (2021), may contribute to the different results regarding latent heating stated on L759–762 of the revised manuscript. However, it would be very difficult to determine how exactly the differences in methodology may contribute to the different results regarding latent heating.

Line 716-720: Does the results are consistent when the metric of forecast skill replaced from RMSE to ACC, which basically represents the pattern difference in synoptic flow?

**Response:**

The identification of low-skill periods and high-skill periods and subsequent comparisons of the Arctic environment and ACs during low-skill periods and high-skill periods are based on using RMSE as the metric to evaluate forecast skill of the synoptic-scale flow over the Arctic. Although ACC is another metric that could be used to evaluate forecast skill of the synoptic-scale flow over the Arctic, using ACC would require redoing all analyses in the manuscript, which we feel would cost a lot of time. Therefore, we cannot answer this question.

Lines 727-: The reviewer feels that the most of sentences are just repeated in Sections 1-3. Thus, the reviewer recommends adding some discussion (e.g., the response to the above reviewer’s questions and other reasons for the high- (low-) skill forecast with low- (high-) skill AC) in this section.

**Response:**

In our responses to some of your previous comments, we indicate where we have added discussion to section 4.

**Other Revisions to Manuscript**

1. The following sentence from section 2c of the original manuscript was modified as follows:

Original manuscript:

“Minimum SLP is used to characterize the intensity of ACs since minimum SLP is provided by the Sprenger et al. (2017) cyclone climatology.”

Revised manuscript (see L277 of the revised manuscript).

“Minimum SLP is used to characterize the intensity of ACs.”

The sentence was modified as such because the Sprenger et al. (2017) cyclone climatology is not used in the revised manuscript.

2. Also, two sentences found at the end of section 4 in the original manuscript were moved to a new paragraph found at the end of the manuscript, and were modified to fit into this new paragraph.

The original sentences were:

“Low-skill ACs during low-skill periods may pose challenges to human activities in the Arctic that may be impacted by potential hazardous weather conditions associated with this category of ACs. Given the aforementioned distinctions between low-skill ACs during low-skill periods and the other skill categories of ACs, and given the challenges low-skill ACs during low-skill periods may pose to human activities in the Arctic, features and processes influencing the evolution and forecast skill of low-skill ACs during low-skill periods are the topic of a future paper.”

The corresponding modified sentences are found on L868–873 of the revised manuscript.