Response to Referee #1:

GENERAL COMMENTS

This paper uses climate model experiments in which regional Arctic sea-ice decline is imposed, combined with analysis of new CMIP6 data, to better understand the dynamical mechanisms by which Arctic sea-ice decline may influence winter haze pollution extremes in China. The main new result reported is that Pacific sector sea-ice loss increases the likelihood and intensity of haze pollution extremes, due to anomalous transient eddy vorticity fluxes amplifying the negative phase of the EU pattern.

Given the substantial impact of haze pollution extremes on public health, this study represents an important contribution to this research area. The variety of methods used – including targeted single model experiments, new CMIP6 multi-model data, and a variety of interesting diagnostics – is also good. The paper is generally well presented with a good quality and number of figures, and only minor alterations are required to the wording and structure.

However, while this study reports potentially very interesting and impactful results, I am concerned about the statistical robustness of some of the conclusions and, therefore, that the length of the simulations (30 years) may be too short. My specific comments below explain these concerns in detail, which I would like to see addressed.

Response: Thank you very much for the constructive comments and suggestions. We understand your concerns about the robustness of the modeling results in the manuscript. Therefore, we have conducted additional statistical significance tests to demonstrate that these results are robust. We also revised the manuscript to address your other concerns. Please see below our responses (in blue) to your specific comments.

SPECIFIC COMMENTS

Page 1, line 19: I found the use of the word ‘event’ a bit confusing in this study, as the extremes analysed are monthly extremes and ‘event’ – to me anyway – implies a shorter timescale (daily or weekly). It would be helpful to clarify somewhere what is meant by the term ‘event’ here, or to avoid using the term.

Response: Thank you for the suggestion. Since there are many different types of climate extreme events such as cold extremes, heatwaves, droughts, and extreme precipitation, etc., we want to emphasize here that pollution-related air stagnation extremes are the major focus of this study. To avoid possible confusion with time scale-related interpretation, we rephrased the expression here to “monthly air stagnation extremes” and revised all similar expressions throughout the manuscript.

Introduction: This paragraph is far too long, which made the structure of the introduction – which while good – a bit hard to follow. Breaking this up into a few paragraphs would help. The same goes for similarly long paragraphs in other parts of the paper (e.g. page 4 lines 9-40; page 9).

Response: We have followed the suggestion to break those long paragraphs into shorter ones on page 2, page 4, and page 9. Please see the revised manuscript for details.

Page 2, lines 32-35: This sentence is a bit misleading, as it implies that there is a scientific

consensus that high-latitude climate change influences mid-latitude circulation and weather, when there is not (e.g. https://www.nature.com/articles/s41558-019-0662-y). There is lots of evidence suggesting that Arctic sea-ice loss can have an influence on mid-latitudes, but whether it has in the past or will in the future is more unclear (https://onlinelibrary.wiley.com/doi/full/10.1002/wcc.337). Would be good to rephrase the sentence to reflect this (e.g. 'Given the increasing evidence that climate change – especially that occurring in high-latitude regions – may have an influence on middlelatitude circulation').

Response: We agree that there are lots of discussion and ongoing debates on this topic. Knowledge gaps regarding complex interactions between high-latitude and mid-latitudes and physical pathways behind these phenomena still exist. A few climate modeling studies have been conducted to narrow down the uncertainty associated with the influence of high-latitude climate change on mid-latitude weather extremes. Our study was motivated and inspired by these discussions and modeling efforts. To clarify on the current research status, we have rephrased the text as suggested and added more specific discussion and references in lines 2-5 of page 3: "Several possible dynamic pathways linking Arctic warming to midlatitude weather extremes have been proposed and investigated in the past few years (Barnes and Screen, 2015; Overland et al., 2016). However, the observational data and modeling results are sometimes contradictory and are open to different interpretations (Cohen et al., 2020)".

Section 2.1: I found this section jumped around a bit in terms of the definitions of the EU pattern and index, the MCA_Z500 pattern, and the PPI. If possible, could this be restructured so that the definition of each is closer to where it is originally introduced?

Response: Thank you for the suggestion. We revised this section to more clearly describe all the indices used in the manuscript. Please see Section 2.1 in the revised manuscript for details.

Page 3, lines 29-30: It would be helpful to properly explain and define the WSI and ATGI.

Response: The two indices are defined and explained in lines 34-37 of page 3: "WSI was standardized by subtracting time-averaged climatological mean of near-surface wind speed over the 1981-2010 period from the monthly values at each grid cell and then dividing by its standard deviations in the same period. ATGI was the standardized potential temperature gradient field between 925 and 1000 hPa using the same method. These two indices are used to reflect horizontal and vertical dispersions of near-surface air pollutants, respectively."

Page 4, lines 3-5: Do you have a citation for this?

Response: This statement is based on the similarity between MCA_Z500 and EU as well as other teleconnection patterns such as the East Atlantic (EA) pattern (https://www.cpc.ncep.noaa.gov/data/teledoc/ea_map.shtml) and the East Atlantic/Western Russia (EA/WR) pattern (https://www.cpc.ncep.noaa.gov/data/teledoc/eawruss.shtml) over East Asia in winter (e.g., January patterns), as shown in Fig. R1.

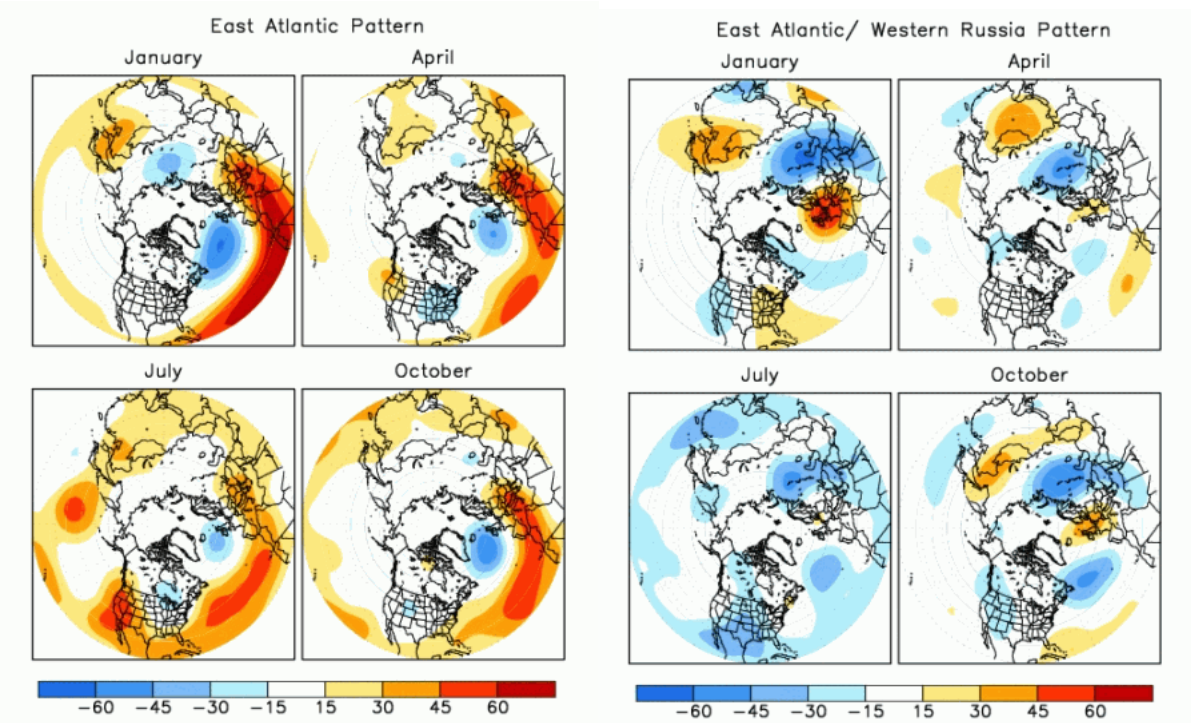

Figure R1: 500 hPa geopotential height anomalies (unit: m) of the East Atlantic pattern (left) and the East Atlantic/Western Russia pattern (right) in different months. These plots are adapted from the NOAA Climate Prediction Center (CPC) website (https://www.cpc.ncep.noaa.gov/data/teledoc/telecontents.shtml; last access: 14 February 2020).

The major difference between EU and other planetary-scale teleconnection patterns is in the wave propagation pathways in the upstream regions such as the Atlantic and Europe, while they share similar configurations in the downstream regions over East Asia. All these teleconnection patterns can be excited by either internal variability or localized forcings (Simmons et al., 1983; Sardeshmukh et al., 1988; Liu et al., 2014; Lim, 2015). To clarify this, we added examples and references in lines 15-18 of page 4 in the revised manuscript as:
"However, it's worth noting that this regional MCA_Z500 pattern can also be excited by other large-scale teleconnection processes such as the East Atlantic pattern or the East Atlantic/Western Russia pattern associated with both natural variability and perturbed Rossby wave activity (Lim, 2015; Simmons et al., 1983)."

Section 2.2: Are you able to justify using simulations of only 30 years in length? To me this seems rather short, especially considering my comments regarding statistical robustness below. Indeed, Screen et al. 2014 show that the simulated circulation response to sea-ice loss is small compared to internal variability (i.e. there is a low signal-to-noise ratio), and specifically that at least 70 year-long experiments are required to simulate a robust mid-tropospheric response to sea-ice loss (https://link.springer.com/article/10.1007/s00382-013-1830-9). Similarly, simulations submitted to PAMIP (the Polar Amplification Model Intercomparison Project) are required to be at least 100 years long due to this low signal-to-noise ratio (https://www.geosci-model-dev.net/12/1139/2019/). Also, many studies using WACCM to investigate the response to sea-ice loss use longer simulations (e.g. England et al. 2019 use 151 years, https://journals.ametsoc.org/doi/full/10.1175/JCLI-D-17-0666.1; Sun et al. 2015 use 161 years,

https://journals.ametsoc.org/doi/full/10.1175/JCLI-D-15-0169.1; Zhang et al. 2018 use 60 years, https://advances.sciencemag.org/content/4/7/eaat6025).

Response: Thank you for the comment and references. Several studies, including those in your comment, have indicated that the signal-to-noise ratio associated with the Arctic influence on midlatitude weather is lower than internal variability, which motivated the long-term simulations in those studies to try to isolate a robust atmospheric response in the middle latitudes to Arctic sea ice loss and Arctic amplification. However, most, if not all, Arctic-midlatitude impact studies focused on the response in ensemble seasonal mean state rather than monthly extreme values in our case. We want to emphasize that the modeling responses could be very different in terms of these two metrics. This is evident by comparing the changes in ensemble mean values (Table S2 in the Supplement) with those in extreme values (Table S3/S4 in the Supplement) of each sensitivity experiment. It can also be clearly demonstrated by the following conceptual changes in temperature distribution and their effects on extreme values (Fig. R2). In this IPCC report (2012), three distinct distribution changes in response to climate change have been proposed: shifted mean, increased variability, and changed symmetry, which suggest complex relationship between changes in ensemble mean and extreme values. We followed this analysis framework to characterize modeling responses in our climate sensitivity experiments and found the SENSr2 results of interest fall into the "Changed Symmetry" category (as shown in Fig. S2/S3 in the Supplement).

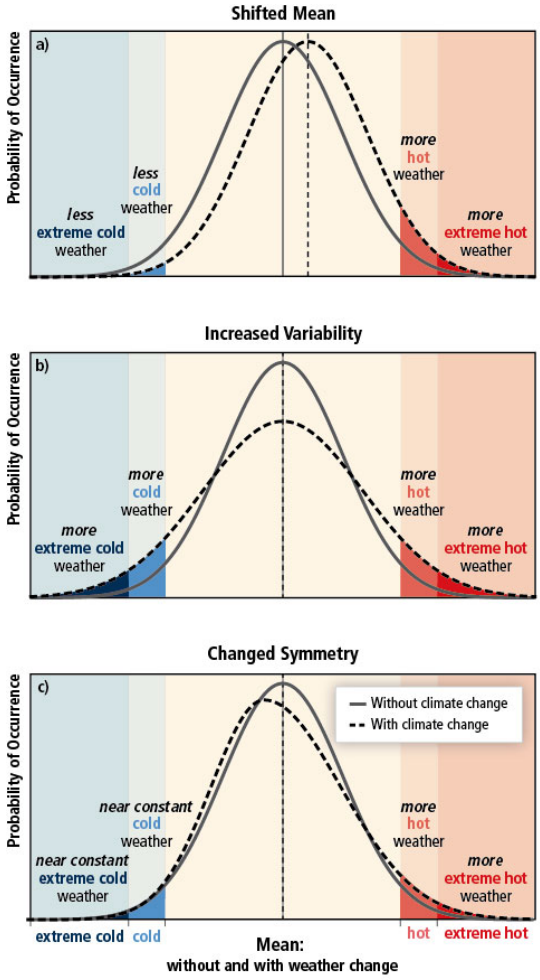

Figure R2: The effect of distribution changes on temperature extremes. Different changes in temperature distributions between present and future climate and their effects on extreme values of the distributions: (a) effects of a simple shift of the entire distribution toward a warmer climate; (b) effects of an increase in temperature variability with no shift in the mean; (c) effects of an altered shape of the distribution, in this example a change in asymmetry toward the hotter part of the distribution. This plot is adopted from Figure SPM. 3 in IPCC (2012).

To evaluate statistical significance of the changes in positive extreme probability, we repeatedly resample a subset of modeling years in SENSr2 for 10,000 times and then use a non-parametric kernel density estimation (KDE) function to re-estimate the probability of ECP_PPI (Fig. R3) positive extremes in each subset comparing with their CTRL counterpart. We try two different methods of resampling: without replacement and with replacement for multiple subset sizes (10, 15, 20, 25, 30). Sampling without replacement does not allow duplicated modeling years while sampling with replacement generates much more combinations and larger variances of subsets. Since there are numerous combinations of resampled subsets (except the subset size of 30 years without replacement, which has only one unique combination of all data), we plot the empirical probability density distributions of positive extreme probabilities using kernel density estimation for each subset size (similar to Fig. 8 in Screen et al., 2014). Please note that the actual sample size in each subset should be multiplied by 3 because we use monthly data in winter (Dec-Jan-Feb) rather than seasonal average to detect climate extremes. For example, the total sampling size for the subset of 10 years is 3 months × 10 years = 30 months. The autocorrelation among these winter months is low and insignificant (please see the response to the next question for details).

After obtaining these empirical PDFs, we estimate the corresponding value of the CTRL positive extreme probability in these PDFs to test the hypothesis that the SENSr2 positive extreme probability is significantly larger than the CTRL run (the CTRL positive extreme probability is always 0.05 since the 95^{th} percentile of CTRL data is chosen as the positive extreme threshold). As shown in Fig. R3 below, the chance of SENSr2 ECP_PPI positive extreme probability $\leq$ CTRL ECP_PPI positive extreme probability (0.05) is about 3% when the subset size exceeds 15 years (45 months) without replacement (Fig. R3a), or when the subset size exceeds 25 years (75 months) with replacement (Fig. R3b). Another way to demonstrate this is to plot the positive extreme probability estimates and their 95% percentile ranges against different ensemble sizes (Fig. R4; we updated Fig. S3/S4 in the supplement using the same method here), which suggest the same conclusion. The ensemble averaged estimates of the SENSr2 ECP_PPI positive extreme probability are also quite similar among different ensemble sizes (~0.11) and more than double of the CTRL positive extreme probability (0.05). Therefore, we are confident that the current modeling simulation length of 30 years is long enough to detect significant extreme probability changes, which is the primary research objective of this study.

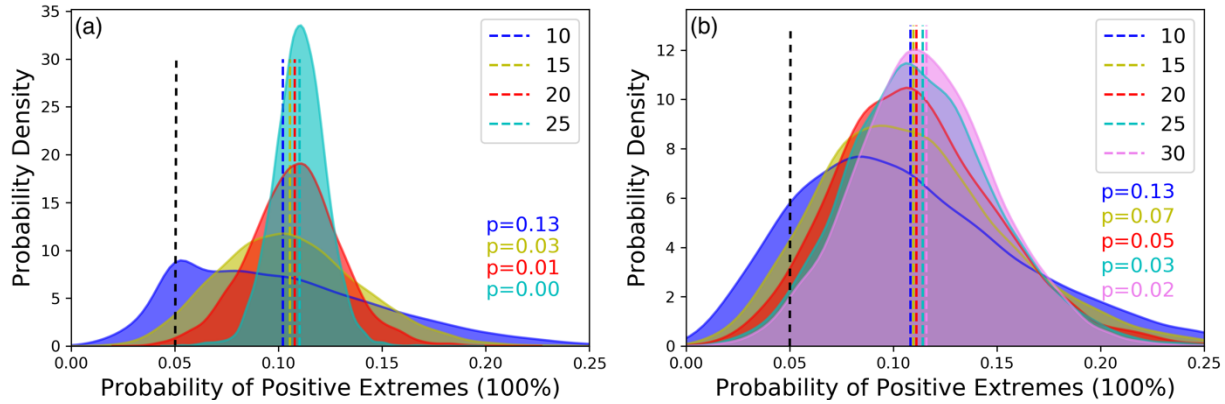

Figure R3: KDE-based probability estimates of ECP_PPI positive extremes in SENSr2 based on different ensemble sizes of subsets (a) without and (b) with replacement in bootstrap resampling (n=10,000). The p values on bottom-right corners are the probabilities of 0.05 (the CTRL positive extreme probability shown as the black dash line) in each PDF curve. The colored dash lines are ensemble averaged probabilities of SENSr2 positive extremes for each subset size. Note that no PDF curve is available for nsize=30 without replacement in (a) because of the uniqueness of the sampling combination.

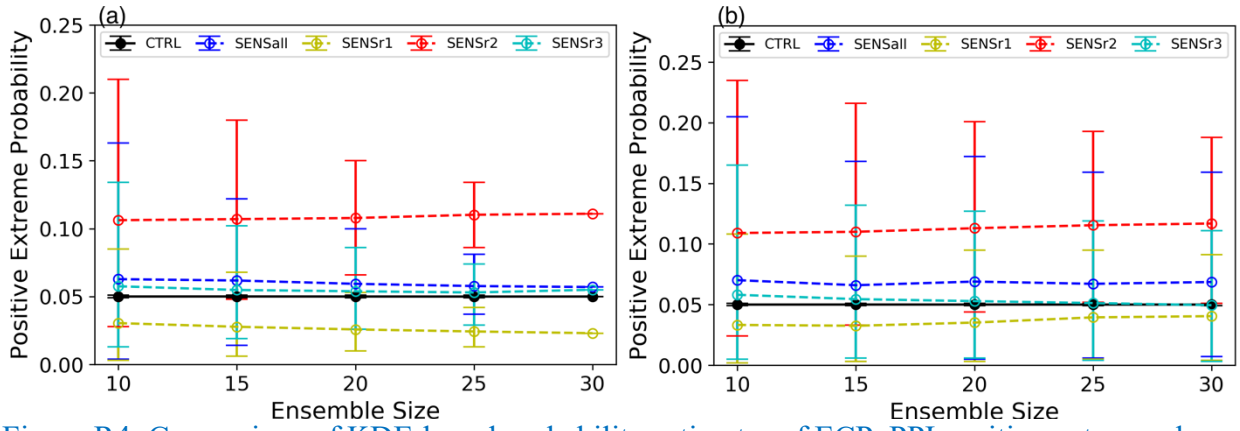

Figure R4: Comparison of KDE-based probability estimates of ECP_PPI positive extremes based on different ensemble sizes of subsets (a) without and (b) with replacement in bootstrap resampling (n=10,000). The error bars denote the 95% percentile range (2.5% to 97.5%) for positive extreme probability values at each ensemble size.

Page 6, lines 4-6: You say that you have 90 samples when conducting this statistical test, and so I presume you are assuming 90 degrees of freedom. However, have you checked whether the MCA_Z500 and/or ECP_PPI indices are autocorrelated (e.g. between consecutive months or lag-1), and therefore whether 90 degrees of freedom is an overestimate?

Response: Thank you for the suggestion. We can treat each sequence of monthly data in Dec, Jan, and Feb as three sampling groups. The essence of this question is whether two consecutive groups of monthly data are independent or not. We test the lag-1 relationship in both MCA_Z500 and ECP_PPI indices by checking the Pearson correlation coefficients between two consecutive monthly groups. If they are not independent from each other, then we would expect statistically significant correlations between these paired groups. Table R1 and Table R2 show the correlation coefficients for both indices and their corresponding two-tailed p-values,

respectively, suggesting insignificant correlations in most MCA_Z500 pairs and all ECP_PPI pairs.

Table R1: Correlation coefficients of the MCA_Z500 and ECP_PPI indices between two consecutive months in each modeling experiment

r	MCA_Z500		ECP_PPI	
	Dec-Jan	Jan-Feb	Dec-Jan	Jan-Feb
CTRL	0.27	0.01	-0.02	-0.28
SENSall	0.43	0.05	0.24	0.19
SENSr1	0.003	0.28	-0.04	-0.03
SENSr2	0.17	0.53	0.07	-0.17
SENSr3	0.25	0.03	0.14	-0.23

Table R2: Two-tailed p-value of the MCA_Z500 and ECP_PPI correlation coefficients between two consecutive months in each modeling experiment

p-value	MCA_Z500		ECP_PPI	
	Dec-Jan	Jan-Feb	Dec-Jan	Jan-Feb
CTRL	0.15	0.95	0.90	0.14
SENSall	0.02	0.81	0.20	0.31
SENSr1	0.99	0.14	0.84	0.87
SENSr2	0.37	0.003	0.72	0.37
SENSr3	0.18	0.86	0.47	0.22

Since the Pearson correlation coefficient is highly sensitive to outliers, we also plot the scatter plots based on the winter consecutive monthly MCA_Z500 data in SENSall and SENSr2 that show possible correlations. As shown in the plots, the Pearson correlations are mainly contributed by two MCA_Z500 outliers (in the red circle) on bottom-left corners between December and January in SENSall (Fig. R5a), and one MCA_Z500 outlier (in the red circle) on top-right corner between January and February in SENSr2 (Fig. R6b). After removing these outliers, the correlations would largely decrease to insignificant levels (SENSall: ($r=0.13$, $p=0.52$); SENSr2: ($r=0.36$, $p=0.06$) after removing the outliers in red circles). Another way to show the large impact of outliers on the Pearson correlation coefficients is to use the non-parametric Kendall rank correlation as an alternative, which is more suitable for small sample sizes without the Gaussian distribution assumption. The Kendall rank correlation coefficients for these MCA_Z500 data in Dec-Jan of SENSall and Jan-Feb of SENSr2 are ($r=0.16$, $p=0.23$) and ($r=0.26$, $p=0.04$), respectively. Both are much smaller than the Pearson ones listed in the above tables. Actually, the lifetime of most severe pollution events is shorter than one month, and the memory effect of the atmosphere is also short. Therefore, we feel it's acceptable to treat these monthly data as independent samples and the degree of freedom of 90 is considered a roughly accurate estimate for the statistical tests in the manuscript.

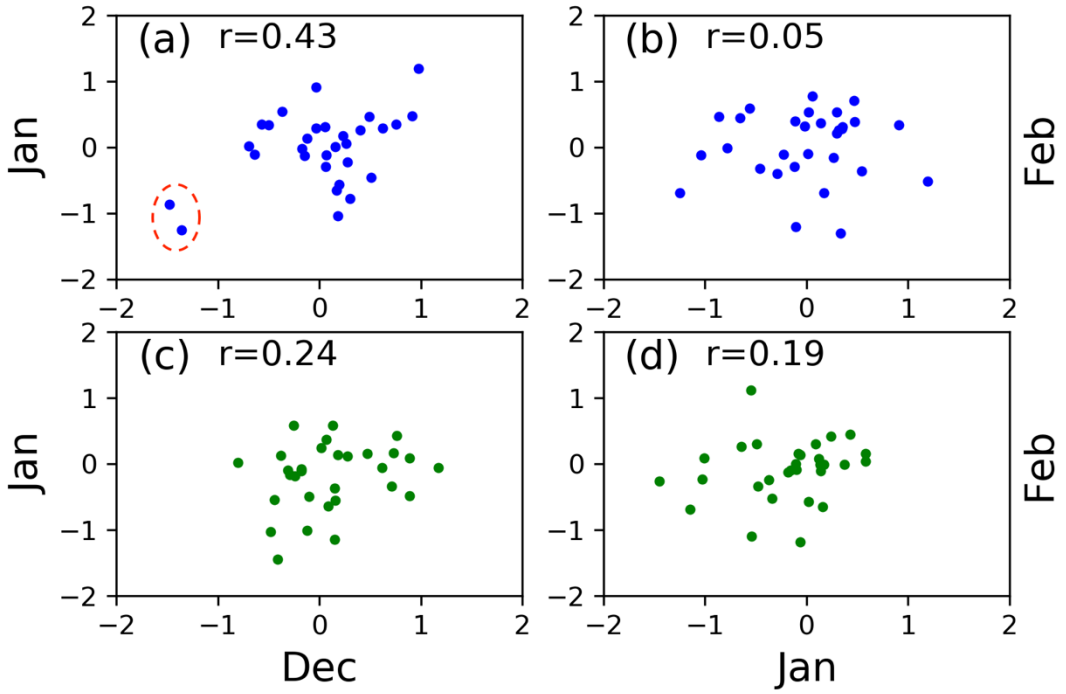

Figure R5: Scatter plots for the MCA_Z500 and ECP_PPI indices in consecutive months of the SENSall experiment. (a) the paired MCA_Z500 indices in December and January; (b) the paired MCA_Z500 indices in January and February; (c) the paired ECP_PPI indices in December and January; (d) the paired ECP_PPI indices in January and February. The red circle in (a) shows the outliers contribute largely to the Pearson correlation coefficient.

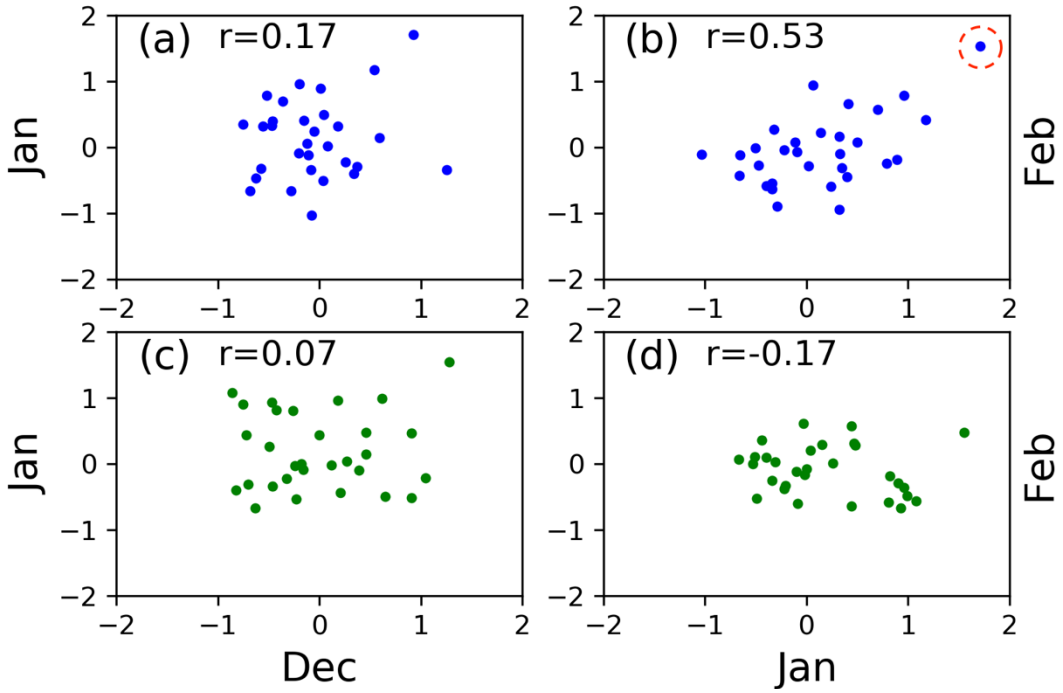

Figure R6: Scatter plots for the MCA_Z500 and ECP_PPI indices in consecutive months of the SENSr2 experiment. (a) the paired MCA_Z500 indices in December and January; (b) the paired MCA_Z500 indices in January and February; (c) the paired ECP_PPI indices in December and January; (d) the paired ECP_PPI indices in January and February. The red circle in (b) shows the outlier contributes largely to the Pearson correlation coefficient.

Page 6, lines 36-39: Relating to the above comment, did you account for autocorrelation when conducting this bootstrapping method (e.g. as done in your previous paper using the moving blocks method https://advances.sciencemag.org/content/3/3/e1602751)? If there is autocorrelation it may be that the uncertainties given by the bootstrap method (Tables S3 and S4) may be underestimated, and therefore the statistical robustness of the differences between the perturbation and CTRL experiments overestimated.

Response: As shown in our response to the previous comment, the autocorrelation in monthly data is negligible in most cases. Therefore, we used the standard bootstrapping method in this manuscript. The reason we used the moving block bootstrap method in our previous study (Zou et al., 2017) is that we used daily data in that study. The autocorrelation problem in these daily time series is much more severe than the monthly data used in this study, so the moving block bootstrap method was applied there. For the monthly data with less concern about autocorrelation, the standard bootstrap method was applied in the previous study (Zou et al., 2017) that is the same with the practice here.

Page 8, lines 21-23: It should be noted that these correlations are not statistically significant at most grid points (but perhaps the correlation would be significant if you used an area average?).

Response: This figure helps to identify the Arctic sub-regions with potential influence on the atmospheric teleconnection as well as regional ventilation in China. If averaging SIC over those R2 areas with positive correlations with the EU index, the regional averaged SIC-EU correlation coefficient is r=0.38 (p=0.02), which is statistically significant at the 0.05 significance level. We added this regional averaged correlation to lines 3-4 of page 9 in the revised manuscript.

Page 9, lines 20-23; Tables S3 and S4: Can you justify why you use the standard deviation here? The numbers in these tables for the SENSr2 experiment contain one of key results of this paper, suggesting that there is an increase in the likelihood and intensity of MCA_Z500 and ECP_PPI positive extremes in response to sea-ice loss in the R2 region. However, by using just the standard deviation it maybe cannot be said that the extremes in SENSr2 are significantly different statistically from those in CTRL.
I may be wrong, but a 95% confidence interval seems more appropriate to test whether the difference is statistically robust? Since a 95% confidence interval will be larger, the 9% 3% figure in Table S3 for SENSr2 MCA_Z500 may not actually be significantly different from CTRL (5% 0%).

Response: Thank you for the suggestion. We redid the bootstrap analysis with replacement for 10,000 times to estimate the 95% percentile range for all the indices listed in Table S3 and Table S4. Please see below the updated tables (we included here for your convenience):

Table S3. The bootstrap (nboot=10000) estimates (ensemble mean and 95% percentile range) of positive extreme probabilities of the MCA_Z500 and ECP_PPI indices in the WACCM experiments

	CTRL	SENSall	SENSr1	SENSr2	SENSr3
MCA_Z500	5.0%	3.7% (0-13.5%)	3.3% (0-9.2%)	7.5% (0.8-16.4%)	4.1% (0-12.8%)
ECP_PPI	5.0%	7.0% (0.7-16.1%)	4.1% (0.4-9.2%)	11.6% (5.2-18.4%)	5.0% (0.2-11.0%)

Table S4. The bootstrap (nboot=10000) estimates (ensemble mean and 95% percentile range) of positive extreme intensities of the MCA_Z500 and ECP_PPI indices in the WACCM experiments

	CTRL	SENSall	SENSr1	SENSr2	SENSr3
MCA_Z500	1.14 (0.75-1.72)	1.00 (0.77-1.35)	1.07 (0.81-1.44)	1.27 (0.90-1.68)	1.03 (0.77-1.41)
ECP_PPI	0.86 (0.63-1.40)	0.91 (0.70-1.25)	0.94 (0.72-1.31)	1.12 (0.90-1.42)	0.84 (0.66-1.13)

The new estimates don't change our conclusion in the manuscript, suggesting significantly increased probability and intensity of ECP_PPI in SENSr2. This is also evident in Fig. R3b of the previous response. The increase in MCA_Z500 is less significant than that in ECP_PPI, which might be attributed to the smaller signal-to-noise ratio in large-scale dynamic processes. Extended climate sensitivity experiments could be conducted in the future to evaluate the robustness of these large-scale dynamic responses.

Page 9, line 23 to page 10, line 28: Results relating to changes in the ensemble mean of the MCA_Z500 and ECP_PPI indices are presented and discussed as if they are statistically robust (e.g. 'The differences in the MCA_Z500 and ECP_PPI responses among the four sensitivity experiments in extreme members and ensemble means also suggest complex relationships between Arctic sea ice loss and mid-latitude weather changes'). However, they are only statistically significant for SENSr1 ECP_PPI (p=0.04) – see Table S2. These paragraphs should be edited so that is clear whether the results being presented and discussed are robust or not.

Response: We rewrote the paragraphs to clearly indicate the robustness of changes in both ensemble mean and extreme values of both indices. Please see below the updated paragraphs in Section 3.2 of the revised manuscript.

"To examine the regional circulation and ventilation responses to these changes in the high latitudes, we fit the CDF and PDF curves of MCA_Z500 and ECP_PPI based on CTRL and SENS monthly results in winter. Figure 3 shows the CDF changes of simulated MCA_Z500 (Fig. 3a) and ECP_PPI indices (Fig. 3b) between sensitivity and CTRL experiments. It is clear that both indices show more significant changes in their extreme members than in medians or ensemble means, especially in SENSr2 driven by SIC and SST changes in the Pacific sector of the Arctic (R2 in Fig. 1b). In SENSr2, the occurrence probability of MCA_Z500 positive extremes increases by 50% from 5.0 to 7.5% (95th percentile range: 0.8-16.4%) (Fig. 3a; Table S3 in the Supplement), while the ECP_PPI positive extremes increases by 132% to 11.6% (95% percentile range: 5.2-18.4%) (Fig. 3b; Table S3 in the Supplement). Meanwhile, the intensity of positive extreme values of the two indices also increases by 11% and 30%, respectively (Table S4 in the Supplement). The increase in the teleconnection pattern index MCA_Z500 is less significant than that in the regional air stagnation index ECP_PPI, suggesting a potential

nonlinear relationship between large-scale circulation and regional stagnation. Only SENSr2 shows statistically significant increases of ECP_PPI in terms of positive extreme probability and intensity, and the significance of such increases is independent from the fitting method being used (i.e., still valid with nonparametric curve fitting). The substantially increased positive extremes in SENSr2 contribute to the positive responses in its ensemble mean, making SENSr2 the only sensitivity experiment with positive ensemble mean ECP_PPI (0.03, not statistically significant). In comparison, other SENS experiments generally show negative ensemble mean ECP_PPI values due to left-shifted CDF curves at most percentiles. For instance, SENSr1 is the only experiment showing robustly decreased ECP_PPI at all percentiles in its CDF curve (Fig. 3b), contributing to its negative ensemble mean of ECP_PPI (-0.13) that is statistically significant at the 0.05 significance level (Table S2 in the supplement). This result implies an overall improvement of the ECP regional ventilation driven by the SIC and SST changes in the Barents-Kara Seas (R1 in Fig. 1b), while the ventilation responses are more random driven by sea ice loss in other Arctic regions."

Section 3.4: Why has only the ECP_PPI index been calculated for the CMIP6 results, and not the MCA_Z500 index, when both were for the WACCM results? This seems quite key, since it is MCA_Z500 that demonstrates a dynamical (and therefore more causal) connection between sea-ice loss and ECP_PPI.

Response: Thank you for the suggestion. We have now added the time series and changes of MCA_Z500, based on the reanalysis and CMIP6 results, in a new supplementary Figure S8 (shown as Fig. R7 here). The MCA_Z500 projection results also show right-shifted positive extremes in future time periods, with the largest shift emerging during the P3 period in concurrence with the strongest decline of Arctic sea ice. Please see Section 3.4 in the revised manuscript for details.

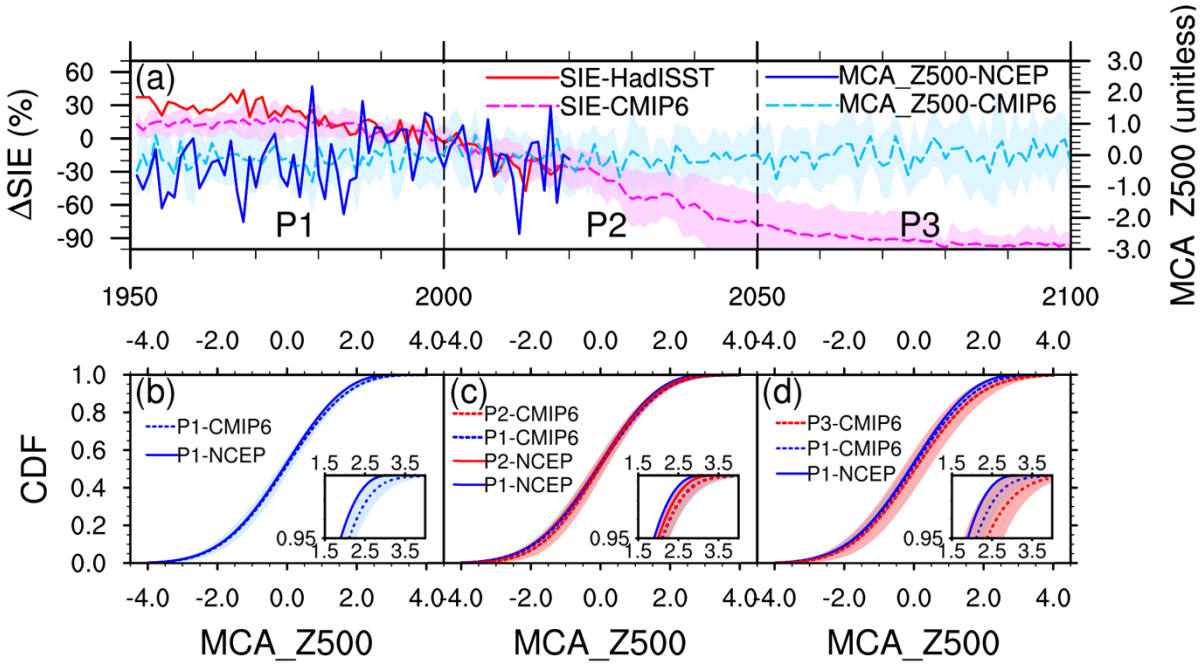

Figure R7. Historical simulations and future projections (under the SSP5-8.5 scenario) of Arctic sea ice and regional circulation in observational and reanalysis data and CMIP6 models. (a) time

series of the Arctic SIE relative changes (unit: %; relative to 1981-2010) in preceding September and MCA_Z500 (unitless) in DJF of the following winter (using years of January for X-axis labeling). The solid lines denote observation- and reanalysis-based Arctic SIE and MCA_Z500 from 1950 to 2019. The dashed lines denote ensemble mean and the color shading denotes ±1 standard deviation of the 8 CMIP6 models (see Table S1 for model details) from 1950 to 2100. Note that the SIE time series were shifted forward by one year to align with the MCA_Z500 data; (b) comparison of MCA_Z500 CDF curves between the NCEP reanalysis data and the CMIP6 models in the P1 time period from 1951 to 2000. The inset denotes the distributions of positive extremes ($\geq MCA\_Z500_{\mathrm{P1}}^{95^{th}}$). The color shading denotes ±1 standard deviations in the 8 CMIP6 models; (c) Same as (b) but for the comparison between P1 and P2 (2001-2050) time periods as well as between the NCEP reanalysis data and the CMIP6 models; (d) same as (b) but for the comparison between P1 and P3 (2051-2100) time periods as well as between the NCEP reanalysis data and the CMIP6 models.

Figure 1, Figure S1, Figure 5 (a) and (c): It would be useful to indicate in the captions that these plots are for observational/reanalysis data, rather than for the sensitivity experiments conducted. For Figure 5 (a) and (c) specifically this is mentioned initially, but it would be clearer to say this in the caption after (a) and (c) as well.

Response: We add the descriptions in the figure captions as suggested.

Figure 3: In the caption it says 'Atmospheric circulation and regional air stagnation responses to the Arctic sea ice forcing in the WACCM experiments'. However, what is in the figure is the absolute CDFs for the CTRL and SENS experiments, rather than differences between the SENS experiments and CTRL (what is normally defined as the 'response'). The use of 'response' in the caption is therefore confusing and should be changed.

Response: We change the description here to "Comparison of the statistical distributions of atmospheric circulation and regional air stagnation indices in the WACCM climate sensitivity experiments" for clarification.

Figure 4: Since these plots show the difference between the SENSr2 extreme members and the CTRL ensemble mean, rather than the CTRL extreme members, these plots do not just show the effect of the sea-ice forcing imposed, but the combined effect of sea-ice loss and internal variability (which causes extreme events without the need for sea-ice loss). The start of the caption ('Winter atmospheric response to the autumn and early winter sea ice change : : :') should therefore be re-phrased. Also - presumably 'winter' means the 'winter mean' here?

Response: Thank you for the suggestion. We rephrase the Fig.4 caption to "Atmospheric anomalies in WACCM SENSr2 extreme members with respect to the CTRL ensemble mean". These extreme members spread in different winter months. Here the anomalies are based on the differences between the average of these extreme members and the CTRL average.
In the dynamic diagnosis part, we attempt to answer the following two questions:

(1) How does severe air stagnation occur in these SENSr2 extreme members?
(2) Why are there more and intensified air stagnation extremes in SENSr2?

As indicated by your comments, the extreme weather in these ensemble members could result from interactions between atmospheric internal variability and Arctic sea ice forcing. And we do

find constructive interference between sea ice-related anomalous wave activity and the background flow (Fig. S7 in the supplement). Therefore, we use Fig. 4 in combination with the following figures (Fig. 5/6 in the revised manuscript) to answer the first question, and then use Fig. 5/6 and Fig. S7 in the supplement to answer the second question. Please see Section 3.3 of the revised manuscript for detailed analysis.

Figure 5: Why is there stippling to show statistical significance in all figures except this one?

Response: We didn't add stippling to this figure in the previous version because it already has 3 layers (shading, contour, and vectors). Adding stipples would further increase its complexity. In the revised manuscript, we update Fig. 5 in the manuscript by separating the SENSr2 extreme members from their CTRL counterparts and adding stipples for significance tests in each subplot as suggested (shown as Fig. R8 here). A new figure is also added in the supplement (Fig. S7) to isolate the difference between the SENSr2 extreme members and their CTRL counterparts directly ($SENSr2_{extreme}$-$CTRL_{counterpart}$).

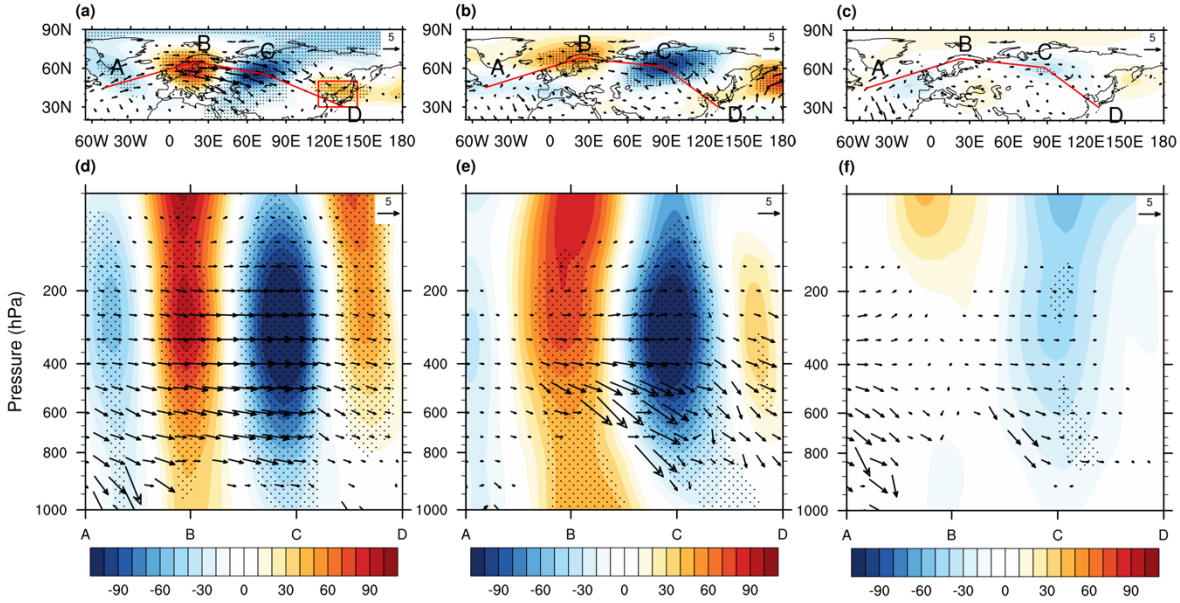

Figure R8: Comparison of atmospheric anomalies in the NCEP reanalysis data and WACCM experiments. (a) reanalysis-based ensemble mean geopotential heights at 500 hPa (color shading, m) and wave activity flux (WAF) at 250 hPa (vectors, $m^2\ s^{-2}$) of the 30 strongest negative EU months in winter (DJF) of 1951-2019 (relative to 1981-2010 climatology); (b) same as (a) but based on the SENSr2 extreme members (relative to CTRL ensemble mean); (c) same as (b) but based on the CTRL counterparts of the SENSr2 extreme members (relative to CTRL ensemble mean); (d) reanalysis-based vertical cross section of geopotential heights (color shading, m) and WAF (vectors, $m^2\ s^{-2}$) of the ensemble mean negative EU months along the wave propagation path shown in (a); (e) same as (d) but based on the SENSr2 extreme members (relative to CTRL ensemble mean); (f) same as (e) but based on the CTRL counterparts of the SENSr2 extreme members (relative to CTRL ensemble mean). Note that the vertical components of WAF in (c)-(d) were scaled up by 200 for clear illustration. The stipples denote the 0.05 significance level.

Figure S4: 'Relative changes' to what?

Response: Here the "relative changes" are changes in terms of percentages rather than absolute values. These percentages are calculated based on the relative concentration differences in SENS extreme members using the CTRL ensemble mean concentration as benchmark. For clarification, we rephase the Fig. S4 caption to "Spatial distributions of surface $PM_{2.5}$ concentration percentage changes (unit: 100%) in extreme members of each sensitivity experiment relative to the CTRL ensemble mean result". Fig. S4c is used for direct comparison with Fig. S5 to demonstrate the effectiveness of PPI.

TECHNICAL CORRECTIONS

Page 3, line 23: Perhaps refer to 'Fig 1 (c) and (d)' instead of just 'Fig 1', since not referring to whole figure. If there are similar instances in other parts of the paper, could you perhaps change these too for clarity (e.g. page 4, line 1: 'Fig S1 (b)' rather than just 'Fig S1').

Response: Thank you for the suggestion. We change the references to specific subplots in the revised manuscript.

Page 3, line 33: I'm not sure the definition of PM_10 would be immediately obvious to all readers, although I could be wrong. Perhaps consider including a very brief definition?

Response: The definitions of $PM_{2.5}$ "(particulate matter with aerodynamic diameters of 2.5 micrometers or less)" and PM_{10} "(particulate matter with aerodynamic diameters of 10 micrometers or less)" have been added after its first appearance in line 7 and line 9 of page 2.

Page 6, line 11: 'these' should be 'those'

Response: Thank you. It's changed to "those".

Page 9, line 24: 'of two indices' should be 'of the two indices'

Response: Thank you. It's changed as suggested.

Figures 3 and 7: 'inlet' should be 'inset'

Response: Thank you. All typos have been changed to "inset" in the captions of Fig. 3 and Fig. 7.

Figure 6 (a) and (b): This rainbow colour scale is not colour-blind friendly, so would be hard to interpret for some people. Perhaps use a white to blue scale, with blue indicating stronger winds?

Response: Thank you for the kind reminder. We change the color bar in Fig. 6a/b and line colors in Fig. 3 to be color-blind friendly.

Tables S3 and S4: 'MAC_Z500' in tables should be 'MCA_Z500'

Response: Thank you. The typos have been corrected.

References

IPCC, Field, C.B., Barros, V., Stocker, T.F., Qin, D., Dokken, D.J., Ebi, K.L., Mastrandrea, M.D., Mach, K.J., Plattner, G.-K., Allen, S.K., Tignor, M., and Midgley, P.M. (eds.): Managing the risks of extreme events and disasters to advance climate change adaptation. A special report of working groups I and II of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK, and New York, NY, USA, 582 pp, 2012.

Lim, Y. K.: The East Atlantic/West Russia (EA/WR) teleconnection in the North Atlantic: climate impact and relation to Rossby wave propagation, Clim Dynam, 44, 3211-3222, 2015.

Liu, Y. Y., Wang, L., Zhou, W., and Chen, W.: Three Eurasian teleconnection patterns: spatial structures, temporal variability, and associated winter climate anomalies, Clim Dynam, 42, 2817-2839, 2014.

Sardeshmukh, P. D., and Hoskins, B. J.: The Generation of Global Rotational Flow by Steady Idealized Tropical Divergence, J Atmos Sci, 45, 1228-1251, 1988.

Screen, J.A., Deser, C., Simmonds, I. and Tomas, R.: Atmospheric impacts of Arctic sea-ice loss, 1979–2009: Separating forced change from atmospheric internal variability, Clim. Dynam., 43, 1-2, 333-344, 2014.

Simmons, A. J., Wallace, J. M., and Branstator, G. W.: Barotropic Wave-Propagation and Instability, and Atmospheric Teleconnection Patterns, J Atmos Sci, 40, 1363-1392, 1983.

Zou, Y. F., Wang, Y. H., Zhang, Y. Z., and Koo, J. H.: Arctic sea ice, Eurasia snow, and extreme winter haze in China, Sci Adv, 3, 2017.