Authors response point-by-point to the reviews, a list of all relevant changes made in the manuscript, and a marked-up manuscript version. NB. From the Figures 1-6, the latter is the revised one.

Authors Response to Interactive comments on "Adding value to Extended-range Forecasts in Northern Europe by Statistical Post-processing Using Stratospheric Observations" by Natalia Korhonen et al.

The comments are in Black and the responses in blue.

We thank the reviewers for their thoughtful and constructive comments.

We have done several major changes to the manuscript. First, we have examined the mean AO instead of the minimum AO. In this process the Figure 2 was replotted. Second, we have increased the sample size of forecasts and observations by including all cases in Nov-Feb 1981-2016 (not just the cases after the first Monday in each month). In this process the Figures 2, 4, 5, and 6 were replotted. Third, we have in this author response demonstrated the ZMZW at 60 °N and 10 hPa at the start of the forecast and 1-6 weeks after different stratospheric situations. In addition to these, we have done several editings to the manuscript, to clarify it, according to the comments. Below we respond to the reviewers point-by-point.

Best regards

Natalia Korhonen and co-authors

Anonymous Referee #1 Received and published: 16 September 2019

The authors present a genuinely interesting analysis that contains new methods to improve extended-range forecasts. The great improvement in forecast skill must be useful work. I found the manuscript interesting and believe others would as well, but I have several questions and comments in the current form.

Major comments

1. The authors used 'minimum daily AO index' but it seems that there is no clear justification for the use of 'minimum'. Use of the minimum AO index might have more uncertainty because the value fluctuates with a day. The uncertainty would be reduced if the authors use weekly mean value rather than the 'minimum'. It would be helpful to isolate the significant skill increase from sampling issues. Additionally,

the post-processing revises weekly mean temperature. This is an additional reason to justify why we need 'minimum' AO index rather than weekly mean.

As suggested, we have now used the "mean" instead of "minimum" in the current manuscript.

2. I am not sure the how the QBO can modulate AO index at weekly time scale. The QBO has an average period of ~28 month. The QBO phase tends to prevail for the entire season so how can we connect dynamics between weekly variation of the AO and QBO?

QBO gives only monthly impact on the probability forecast, however, the ZMZW at 60 °N at 10 hPa (indicating the strength of the polar vortex) gives weekly impact as it is using the last 10 days preceding the start of the forecast.

3. The authors suggested that great improvement in forecast skill associated with QBO polar vortex connection. The past studies suggested that the EQBO could modulate polar vortex, which in turn lead the AO. However, EQBO and polar vortex is not much coincided (Fig. 2). The number of EQBO (u wind <10m/s) is 34 and week vortex (ZMZW <3.8m/s) is 9. Sum of them is 43 but the number of SWIneg cases are 41, which means the events satisfying both condition is only 2. This implies that there should be relationship between EQBO and AO, which is independent to polar vortex. The authors should elaborate introduction and discussion for the prediction skill source for the statistical post-processing.

In the current revised manuscript we increased the sample size by including all cases in Nov-Feb 1981-2016 (not just the cases after the first Monday in each month). Thereby the number of cases, n, in Figure 2 is now higher than in the discussion paper. We plotted Figure AR1 (in this document) to demonstrate how the mean ZMZW at 60 °N 10 hPa was at the start of *EQBO*, *WQBO* etc. (Fig. AR1 a) and 1-6 weeks after *EQBO*, *WQBO* etc. (Fig. AR1 b-d). Fig. AR1 a) shows that the weaker than 3.8m/s ZMZW does indeed not often coincide with the *EQBO* etc. at the start of the forecast. However, Fig. AR1 b-d shows that the mean ZMZW at 60 °N 10 hPa is lower 3-6 weeks after *EQBO* than after *WQBO*.

Editings to Introduction:

"Scaife et al. (2014, Fig.4a) demonstrated indicators of a more negative AO in the easterly QBO at level 30 hPa than in the westerly QBO phase at this level."

Editings to Discussion:

"We investigated the prediction source of the QBO at 30 hPa. In line with Scaife et al. (2014) we found that AO was weaker 1-6 weeks after EQBO that WQBO at 30 hPa. As negative AO index enables cold air outbreaks to Northern Europe (Thompson et al. 2002, Tomassini et al. 2012) and positive AO index tends to bring milder and wetter than average weather to Northern Europe (Limpasuvan et al. 2005), we tested the predictor SWI_{neg}/SWI_{plain} as a predictor of mean surface temperature in Northern Europe for forecast weeks 3-6. We found that the mean surface temperature anomalies in Northern Europe in

November–February in 1981–2016 after SWIneg and SWIplain were statistically significantly different, with anomalously cold surface temperatures more common 3–6 weeks after SWIneg."

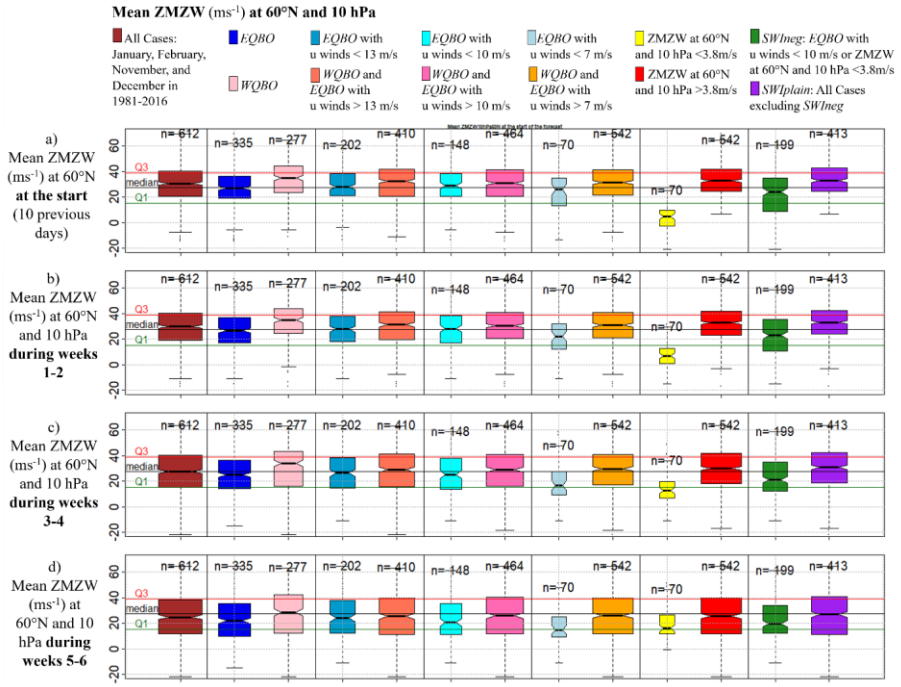


Figure AR1. Mean Zonal Mean Zonal Wind (ZMZW) at 60N and 10 hPa. The red, black and green vertical lines represent the third quantile, the median, and the first quantile, respectively, of the ZMZW at 60 °N and 10 hPa in November-March 1981-2016. The horizontal line dividing each box into two parts shows the median of each data, the ends of the box show the lower and upper quartiles, and the whiskers represent the highest and the lowest values excluding outliers. The n written above each box indicates the number of observations in each group. The widths of the boxes have been drawn proportional to the square-roots of n. The notches of each side of the boxes were calculated by R boxplot.stats. If the notches of two plots do not overlap, this is 'strong evidence' that the two medians differ (Chambers et al., 1983, p. 62). ZMZW=zonal mean zonal wind. SWI=Stratospheric Wind Index.

Minor comments

The annotation “Fig. 2 EQBO and vortex ZMZW < 3.8m/s” (green) does not correspond to decision tree in Figure. 3. Please revise it for the better understanding.

We removed the colors from Figure 3 to avoid misunderstanding.

Anonymous Referee #2 Received and published: 16 September 2019

Summary

The authors use a post-processing technique based on stratospheric predictors of the Arctic Oscillation on ECMWF forecasts to try to improve predictive skill of surface temperature at weeks 3-6. Overall the paper addresses relevant scientific questions (given the lack of some stratospheric processes such as the QBO and its teleconnections in the forecast model), but certain aspects of the paper could be clarified and reorganized. I also have questions about their particular technique of only using forecasts made on the first week of each month, and using stratospheric data at 10 hPa, rather than in the lower stratosphere which is a better indicator of stratosphere-troposphere coupling. I suggest a major revision.

General Comments

1) More could be explained up front about how the QBO teleconnection processes in particular are not well captured in the forecast models (particularly after early winter, as per Garfinkel et al. 2018), and why. In particular, see Line 14-15, page 3: Here it should be specifically mentioned whether the ECMWF S2S model (the version used here) is able to self-generate QBO variability (most S2S models cannot- instead they are initialized with observed QBO winds and degrade relatively quickly away from that towards model climatology). This should be further emphasized (possibly showing something like the forecasted QBO winds in the tropics compared to observed values for each of the initialization dates used here, so it's clear that the model is missing this process).

The version of the ECMWF used here is IFS cycle 43r1. The representation of the QBO in this version of the IFS is described in detail in Johnson et al. (2019) and Stockdale et al (2018). The skill of the QBO forecasts decreases substantially after the first 2 months of the forecast. This is shown to be sensitive to the parametrization of the tropical non-orographic gravity wave drag in the model (see also Polichtchouk et al. (2018), Polichtchouk et al. (2017)). The amplitude of the QBO tends to weaken through the forecast. All the forecasts in our study are initialised from ERA-Interim reanalysis: the QBO is well represented in the initial conditions but as noted the amplitude tend to weakens during the 6 weeks of the forecast. This is shown in Garfinkel et al (2018) and we have revised the text to clarify this.

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S. P. E., Mogensen, K., Zuo, H., and Monge-Sanz, B. M.: SEAS5: the new ECMWF seasonal forecast system, Geosci. Model Dev., 12, 1087–1117, https://doi.org/10.5194/gmd-12-1087-2019, 2019.

Polichtchouk, I., et al., 2017: What influences the middle atmosphere circulation in the IFS? ECMWF Technical Memorandum No. 809.

Stockdale, T. et al., 2018. SEAS5 and the future evolution of the long-range forecast system, ECMWF Technical Memoranda n. 835. DOI: 10.21957/z3e92di7y

2) Sensitivity testing to the parameters chosen here could be further provided, as I'm not sure I understand what was the motivation for some of these choices. For example, see Line 21-25, page 5: Why were QBO winds at 30 hPa selected and was the sensitivity to this level tested? What about 40 or 50 hPa? Also, I understand using the 60N 10 hPa metric for stratospheric variability, but it's not a great metric for coupling to the surface. I assume you might see much better results using winds at a lower level in the stratosphere, near 50-100 hPa. See for example Karpechko et al. 2017 (https://doi.org/10.1002/qj.3017). What happens if you use a lower level instead?

QBO winds at 30 hPa were chosen because we found better effect on the mean AO in the coming weeks then by the 40 hPa or 50hPa. Also Scaife et al. (2014) demonstrated in their Fig. 4a that the AO was more negative after easterly QBO at 30 hPa that westerly QBO at the same level. As we are using the mean QBO winds of the previous month as predictors, this means that in the 5-6 weeks forecast the QBO observation is 2 to 2.5 months old. Indeed the 30 hPa values in time precede the 40 hPa and 50 hPa.

Karpechko et al. 2017 found that the conditional probability of having a tropospheric signal after SSW depends on the value of AO at 150 hPa during 0-4 days after the ZMZW at 60 N and 10 hPa had turned easterly (central date=CD). We are not able to directly implement this to our post-processing, as we are not using the CD, but for our post-processing of the probabilistic forecasts we are using the ZMZW at 60 N and 10 hPa being below its November-March 10th percentile (3.8m/s) as an indicator for a weak polar vortex at the start of the forecast and probable a more negative surface AO index in the coming 1-6 weeks.

3) Furthermore, presumably this technique of identifying weak vortex periods (Line 25, page 5) misses quite a few observed SSW events in the stratosphere (because they don't all occur in the last 10 days of the month). Could you say something about this, and why your method is still valuable? It also seemed like the sample size for u<3.8 m/s was extremely small (n=9) for the longer record period and must be even smaller for the shorter 1997-2016 period of the hindcasts. I wondered why it was so small I wondered why it was so small (144 forecasts, *10th percentile climatology = 14 times this should happen). Can much be said for a predictor that only happens 9 times in 36 years? What happens if this threshold is relaxed to include more events? This is another reason why getting daily tropical winds so all forecast initialization dates may be used could be valuable. Along these lines, I'm confused about the number of events shown in Figure 2 for the two rightmost columns.

In the revised manuscript we have increased the sample size by including all cases in Nov-Feb 1981-2016 (not just the cases after the first Monday in each month). Now the sample size for u<3.8 m/s is 70 for the 1981-2016 period.

4) Finally, I'm curious about the choice of predictors in terms of their covariability. How many times when you had EQBO (and EQBO with maximum thresholds) did you also see a weak polar vortex at 60N 10mb? I assume these are correlated/concurrent, and the EQBO merely adds additional samples where

the vortex is weak but not weak enough to meet the 3.8 m/s threshold. It would be interesting to state what the mean value of 60N 10mb zonal winds are during the various EQBO thresholds.

Figure AR1 above in this document shows boxplots of the mean ZMZW at 60 °N 10 hPa at the start of *EQBO*, *WQBO* etc. (Fig. AR1 a) and 1-6 weeks after *EQBO*, *WQBO* etc. (Fig. AR1 b-d). Figure AR1 shows that the mean ZMZW at 60 °N 10 hPa is weaker 3-6 weeks after *EQBO* and *EBOQ* with different thresholds (Fig. AR1 c-d) than at the start of the forecast (Fig. AR1 a).

Specific Comments

1) Line 6, page 4: You don't consider forecasts in the summer here though, so why 52 weeks? Shouldn't it just be Nov-Feb forecasts? (also, because it wasn't clear I assume you mean throughout that you use any runs initialized in Nov-Feb, but the forecasts in Feb obviously forecast into March/April, correct?)

In this part and in Figure 1 we have actually verified reforecasts run for every week (52 weeks) of years 1997-2016 (20 years), i.e., 52*20=reforecasts 1040. The *SWI* post-processing, however, was done only for reforecasts initialized in Nov-Feb. And correct, the forecast initialized in February forecast into March/April and those are also included in the post-processing.

2) Line 19-20, page 4: Are the operational forecasts used in this study? I'm not sure I understand why the CRPS is adjusted for 51 members if only the hindcasts with 11 members are shown throughout, but maybe I missed where the operational forecasts were used.

There are no operational forecasts used in this study. We refer to the operational forecasts of the ECMWF's IFS which reforecasts we verify. We added "of the ECMWF's IFS" after "the operational forecasts".

3) Line 23, page 4; line 2, page 5: Not sure what is meant here by "annual mean"- do you mean the average of weeks1-6 across all years? Or the average across all months to get one value for each year? I'm not sure I follow what is meant in this paragraph. Also, you might present Figure 1 in section 2.1 and 2.2 in relation to what is being discussed to make it clearer.

We edited this part by:

We calculated the annual means of the expected $CRPS_{RF}$ across all weeks (1 to 52) of the years 1997-2016 reforecasts. These annual means were computed separately for lead times of 1 week, 2 weeks, 3 weeks, 4 weeks, 5 weeks, and 6 weeks, here called forecast week 1, forecast week2, forecast week 3, forecast week 4, forecast week 5, and forecast week 6, respectively. Further, the skill scores of the annual mean CPRSs, the annual mean CRPSSs, for each lead time were calculated as follows:

$$CRPSS = 1 - \frac{CRPS_{RF}}{CRPS_{clim}} \qquad (3).$$

4) Line12-13, page 5: Isn't this data MERRA-2 reanalysis? Could you be more specific about what this data is, and how it was derived? Also, if you are using ERA-interim to verify the forecasts- why not also use ERA-interim for daily zonal winds, both in the stratosphere polar vortex and in the tropics for the QBO? Singapore winds may not always represent the zonal-mean tropical winds that drive QBO teleconnections. Presumably the forecast model is initialized using the winds in the reanalysis, correct? And if you need daily winds to be able to look at forecasts other than the first forecast initialized each month, you could easily get this data all from one product, rather than three different products.

Yes, this is MERRA-2 reanalysis, we added this information to the manuscript.

In the current manuscript we increased the sample size by including all cases in Nov-Feb 1981-2016 (not just the cases after the first Monday in each month). For QBO we still used the Singapore winds, we just used the previous months' Singapore winds for every weeks' forecast. For the ZMZW at 60N and 10hPa we always used the most recent, the last 10 days' wind reanalysis data. Hence this lead to variation in *SWI* even within the month with constant QBO.

5) Line 18-30, page 5: I found this a bit hard to follow without any visual explanation, and I wonder if it would be clearer to discuss Figures 2 and 3 in this section instead of later.

To clarify what was done, we edited this part by:

As Scaife et al. (2014) demonstrated a more negative AO in the easterly QBO at 30 hPa compared to the westerly QBO at 30 hPa, we explored the AO index 1–6 weeks after following predictors:

- westerly QBO at 30 hPa, the *WQBO*,
- easterly QBO at 30 hPa, the *EQBO*,
- *EQBO* with the maximum of the monthly mean zonal wind components of the QBO between 70 hPa and 10hPa restricted to $7ms^{-1}$, $10ms^{-1}$, and $13ms^{-1}$,
- the daily ZMZW at 60° N and 10 hPa during the last 10 days of the previous month falling below its overall wintertime (November–March 1981–2016) 10^{th} percentile, corresponding a value of 3.8m/s, indicating a weak polar vortex already at the start of the forecast.

The statistical significance of the difference between the AO index following two different stratospheric situations, e.g., the *EQBO* and the *WQBO*, was determined using a two-sided Student's t-test with the null hypothesis that there is no difference. The most statistically significant predictors for weaker AO indexes observed 1–2 weeks, 3–4 weeks, and 5–6 weeks after these stratospheric situations, were used to define a *SWI* to be SWI_{neg}; otherwise, it was defined as SWI_{plain} for the beginning of each winter month (November–February) in 1981–2016.

6) Line 13, page 7: is CRPSS above zero a reasonable metric of skill? CRPSS near zero but positive surely can't be that useful (some of these values in Figure 1 are less than 0.1).

CRPSS above zero means that this probability forecast is at least better that the climatological forecast (here the climatological forecast is a 30 member ensemble of 1981-2011 weekly mean surface temperature observations).

7) Line19, page7: Why was only the minimum daily AO considered and not the mean? Does the mean not change enough?

As suggested, we have now used the "mean" instead of "minimum" in the current manuscript.

8) Line 25, page 8: I think Fig 4p looks very much like Fig 4j.

Yes, and we edited the text to bring this up.

9) Figure 6: How are panels (a,b) different than Figure 1? (other than being two week averages rather than 1 week).

In Figure 1 the CRPSSs are means of the reforecasts of all weeks of years 1997-2016 (52 weeks, 20 years), whereas in Figure (a,b) only reforecast initialized for November-February 1997-2016 are included.

Technical Edits

1) Line 19, page 1; line 22-23, page 3; possibly other locations: specify that you are referring to the previous months' tropical stratospheric wind observations.

We edited the "observations" to be "conditions" as this includes (in addition to tropical stratospheric wind observations) also the MERRA-2 reanalysis of the zonal mean zonal wind (ZMZW) at 60 °N and 10 hPa.

2) Line 2, page 2: not sure what is meant by "experimented during a one year living lab"

"during a one year living lab" was edited to "by a one year piloting phase"

3) Line 3, page 2: put comma after "production"

Done.

4) Line 24, page 2: maybe instead "other definitions have been used"

Done.

5) Line 10, page 3: I would clarify that this paper looked at S2S hindcasts similar to what you are looking at here

We edited to be:

"Even though some S2S models, including the ECMWF's Integrated Forecasting System (IFS, Vitart, 2014), are already able to reproduce the QBO's effect on the polar vortex, they are still underestimating the effect on surface weather (Garfinkel et al. 2018)."

6) Line 2, page 4: remove the word "scale"

Done.

7) Figure 1- dots seem a little blurry, might make sure it's high enough resolution for final version.

Done.

8) Line 29, page 7: add in "maximum" to "QBO's monthly mean zonal wind components" or it doesn't make much sense

Done.

9) Line 5, page 8: add in "where" between "cases" and "the"

Done.

10) Figure 2- might be nice to put in bold those p-values that are less than 0.05. Also, shouldn't the last column be labeled "EQBO with u winds < 10 m/s OR ZMZW at 60N and 10 hPa >3.8 m/s"?

We labeled the last column as suggested.

11) Line 24, page 9: The way this is written is confusing, do you mean p decreased so significance increased?

Yes, here the "decrease" should have been "increase". This sentence was, however, removed while editing the text.

12) Line 29, page 9: add a “to” after “corresponding”

Done.

Anonymous Referee #3 Received and published: 19 September 2019

This paper describes a post-processing method to improve sub-seasonal forecasts of northern European winter temperatures, based on the state of the stratospheric polar vortex and QBO in the period immediately preceding the forecast. The paper is interesting, topical and clearly explained, but I have some reservations about the method that need to be addressed before the paper is suitable for publication.

Major comments

1. I'm not convinced about the inclusion of the QBO as a predictor of the Arctic Oscillation (AO) in the method. The authors note that the westerly / easterly QBO is associated with a stronger / weaker polar vortex, but the polar vortex is already included as the other predictor. Unless the QBO directly influences the AO independently of the polar vortex, it's hard to see how the QBO can provide additional skill in forecasting the AO. The authors state that the results are more significant when both the polar vortex and the QBO are used as predictors of the AO, presumably referring to the results in figure 2. The method partitions the 144 observed winter months (for NDJF, 1981-2016) into two sets based on a stratospheric precursor criterion (eg the polar vortex winds are either anomalously weak, or not). The aim is to make the two sets as distinct as possible in terms of their AO index values. Figure 2 shows the distribution of the AO values for each pair of sets obtained using various different criteria. The partition based on the polar vortex and the QBO (green and purple boxes) does show marginally lower p-values than the partition based on the polar vortex alone (yellow and red boxes) consistent with the authors' claim. However, the partition based on the polar vortex alone is split 9 months to 135 months, leading to quite a large uncertainty in the mean AO-index value for the set of 9 months. I suspect this is leading to a higher p-value. Were other thresholds for the polar vortex winds tried, other than 3.8m/s? In the figure, the difference between the median values appears larger for the partition based on the polar vortex alone (yellow and red boxes) than for the partition based on the polar vortex and QBO (green and purple boxes). This suggests to me that the QBO isn't obviously adding any skill in discriminating between high AO and low AO winters.

In the current manuscript we have increased the sample size by including all cases in Nov-Feb 1981-2016 (not just the cases after the first Monday in each month). Now the sample size for the partition based on the polar vortex alone is 70 (out of 612) for the 1981-2016 period giving more certainty in the mean AO index values. The difference between the median values are larger for the partition based on the polar vortex alone (yellow and red boxes) than for the partition based on the polar vortex and QBO (green and purple boxes), however the sample size for SWI_{neg} (199) is larger than weak polar vortex alone (70) giving more certainty for using this in post-processing the probabilistic 3-4 and 5-6 weeks mean temperature forecasts.

2. The method is based on partitioning the winter months into i) those with an anomalously weak polar vortex and/or easterly QBO, and ii) all the remaining winter months It's not really obvious how this method was arrived at. Have the authors considered also separating out the set of winter months with an anomalously strong polar vortex? It seems like an obvious thing to try, and may provide additional skill in predicting the AO.

Yes, we tried this also, but so far we were not able to find predictors for the stronger AO that would have also improved the forecasting skills.

Minor comments

p2, line 26: How exactly is the Arctic Oscillation index defined? The authors just say it's based on 1000hPa geopotential height for 20-90N.

To clarify the AO, we added:

"In Northern Europe one of the important indicators of the large-scale weather patterns is the phase of the AO. The AO is a climate pattern characterized by winds circulating counter clockwise around the Arctic at around 55°N latitude."

p4 line 19: "the CRPS_RF of the CRPS_rf" - what does this mean?

the CRPS_RF is the expected CRPS (assuming there were 51 members) of the ECMWF's reforecast, and the CRPS_rf is the CRPS of the ECMWF's reforecast with 11 members. We edited the text to clarify this.

p6 lines 15-22: I didn't entirely follow the method for making anomalies here - if you're just taking the mean of the 7 anomalies based on different years, aren't you going to get the same answer as just using all the years?

Yes, we changed this and the text to use just the mean of all years.

p7 line 19: The method defines the AO value as the lowest value of the daily AO index in different weeks of the forecast. Why was this chosen - is it representative of the northern European temperature in those weeks? The weekly mean AO value would presumably be less noisy.

As suggested, we have now used the "mean AO" instead of "minimum AO" in the current manuscript.

p8 line 6: the zonal mean zonal wind threshold is stated as 4.8m/s here, but 3.8m/s in figure 2.

The threshold was here corrected to 3.8m/s.

Formatted: Justified, Line spacing: 1.5 lines

Adding value to Extended-range Forecasts in Northern Europe by Statistical Post-processing Using Stratospheric Observations

Natalia Korhonen[1,2], Otto Hyvärinen[1], Matti Kämäräinen[1], David S. Richardson[3], Heikki Järvinen[4], Hilppa Gregow[1]

[1]Weather and Climate Change Impact Research, Finnish Meteorological Institute, Helsinki, 00101, Finland
[2]Doctoral Programme in Atmospheric Sciences, University of Helsinki, Finland
[3]ECMWF, Reading, UK
[4]Institute for Atmospheric and Earth System Research/Physics, University of Helsinki, 00014, Finland

Correspondence to: Natalia Korhonen (Natalia.Korhonen@fmi.fi)

Abstract. The skill scores of the Extended-Range Forecasts (ERF) of the European Centre for Medium-Range Weather Forecasts (ECMWF) are still quite modest for the forecast weeks 3–6 in Northern Europe. As there are known stratospheric precursors impacting the surface weather with potential to improve ERFs, we aim to quantify the effect of these predictors and post-process the ERFs with them.

During boreal winter the quasi-biennial oscillation (QBO) affects the stratospheric polar vortex; the easterly (westerly) QBO often coincides with weaker (stronger) than average polar vortex. Consequently, the weaker (stronger) than average stratospheric polar vortex is connected to negative (positive) Arctic Oscillation (AO) and colder (warmer) than average surface temperatures in Northern Europe. We developed a stratospheric wind indicator, *SWI*, based on the previous weeks' ~~months'~~ stratospheric wind conditions ~~observations~~ and the phase of the AO during the following weeks. We demonstrate that there was a statistically significant difference in the observed surface temperature within the 3–6 weeks depending on the *SWI* at the start of the forecast. These temperature anomalies were underestimated by the ECMWF's reforecasts.

When our new *SWI* was applied in post-processing the ECMWF's two-week mean temperature reforecasts for weeks 3–4 and weeks 5–6 in Northern Europe during boreal winter, the skill scores of those weeks were slightly improved. This indicates there is some room to improve the ERFs, if the stratosphere-troposphere links were better captured in the modelling.

1 Introduction

Extended-range forecasts (ERF; lead time up to 46 days) by dynamical models have been developed since the 1990s with the aim to fill the gap between the medium-range weather forecasts and the seasonal forecasts. It is known that ERF skills are still rather modest in forecast weeks 3–6 especially in the Northern latitudes. If the skill of the forecasts improves, ERFs have the potential to become an essential element in climate services e.g., in the form of early warnings of climatic extremes. In an

academic project CLIPS (CLImate services supporting Public mobility and Safety), climatic impact outlooks and early warnings of extremes (CLIPS forecasts) were developed by employing the ERF datasets (Ervasti et al. 2018). The CLIPS forecasts were co-designed with the general public in Finland and experimented by a one year piloting phase~~during a one year living lab~~. As many industries, e.g., energy and food production, as well as users from the general public considered they could use and would benefit from reliable ERFs (Ervasti et al. 2018), development of more skillful ERFs is clearly needed.

The European Centre for Medium-Range Weather Forecasts (ECMWF) has produced ERFs routinely since March 2002 (Vitart 2014). The verification results of the ECMWF model's ERF (Buizza and Leutbecher 2015; Vitart 2014) on a sub-continental and a regional scale (e.g., Monhart et al. 2018) demonstrated predictive skill beyond 2 weeks for temperature reforecasts over Northern Europe. ECMWF uses bias-correction of the mean in their automatic products, removing the mean bias computed from the reforecasts, depending on the time of the year (Buizza and Leutbecher 2015). We consider the bias over Northern Europe not to be dependent only on the time of the year but also on the prevailing weather pattern, and therefore, we aim to explore whether known teleconnections such as the strength of the stratospheric polar vortex, the phase of the Arctic Oscillation (AO), and the phase of the Quasi-Biennial Oscillation (QBO) could be used in improving the forecasts.

The stratospheric polar vortex is an upper-level low-pressure area that forms over both the northern and southern poles during winter due to the growing temperature gradient between the pole and the tropics. Strong westerly winds circulate the polar vortex, isolating the gradually cooling polar cap air. The strength of the northern polar vortex varies from year to year and can be indicated by, e.g., the zonal mean zonal wind (ZMZW) at 60 °N and 10 hPa or polar cap temperatures. The stronger the circumpolar winds and the colder the polar cap temperatures are, the stronger is the polar vortex. Planetary waves from the troposphere disturb the northern stratospheric polar vortex, leading to meandering and weakening of the westerlies and occasionally to reverse, i.e., easterly flow (Schoeberl, 1978). This weakening of the stratospheric polar vortex also leads to warming of the polar cap temperatures, sometimes even > 30–40 K within several days. A warming of this magnitude together with a reversal of the ZMZW at 10 hPa at 60 °N is commonly defined as a major sudden stratospheric warming (SSW), albeit other definitions have also been used~~also occur~~ (Butler et al. 2015).

In Northern Europe one of the important indicators of the large-scale weather patterns is the phase of the AO. The AO is a climate pattern characterized by winds circulating counter clockwise around the Arctic at around 55°N latitude. At the surface, the AO index~~, which is constructed from the daily 1000 hPa geopotential height field over 20°N–90°N against monthly mean values,~~ is affected by the strength of the polar vortex with a time lag of about two to three weeks (Baldwin and Dunkerton 1999). A strong polar vortex is characterized by lower than average surface pressure in the Arctic, positive AO index, and strong westerly winds keeping the cold Arctic air locked in the polar region and bringing milder and wetter than average weather to Northern Europe (Limpasuvan et al. 2005). In contrast, a weak polar vortex is characterized by higher than average surface pressure in the Arctic, negative AO index, and the meandering and/or weakening of the polar jet stream and

tropospheric jet stream enabling cold arctic/polar air outbreaks to Northern Europe (Thompson et al. 2002, Tomassini et al. 2012).

During boreal winters, the strength of the stratospheric polar vortex influences the surface weather in the Northern Hemisphere within weeks or months (Baldwin and Dunkerton 2001, Kidston et al. 2015), hence holding a potential for forecasting in that time scale. When making forecasts based on the strength of the polar vortex, a noteworthy phenomenon is also the QBO, a quasiperiodic oscillation of the equatorial zonal wind between downwards propagating easterlies and westerlies in the tropical stratosphere with a mean period of 28 to 29 months (Baldwin et al. 2001). Holton and Tan (1980) found that during the easterly QBO at level 50 hPa the polar vortex was statistically significantly weaker than during westerly QBO at the same level. Further, Scaife et al. (2014) demonstrated indicators of a more negative AO in the easterly QBO at level 30 hPa than in the westerly QBO phase at this level. There is no precise consensus of the mechanisms of this tropical-extratropical connection, but the most common explanation is that the QBO affects the polar vortex via the Holton Tan effect: During easterly QBO, small amplitude planetary waves are reflected back towards the North Pole weakening the polar vortex (Holton and Tan 1980, 1982; Watson and Gray 2014, Gray et al. 2018). ~~Garfinkel et al. (2018) found by model simulation a weakened stratospheric polar vortex during the easterly QBO phase compared to the westerly phase in early winter (October–December).~~

Challenges related to the realistic modelling of the dynamical stratosphere-troposphere coupling have been adduced by Shepherd et al. (2018) and Polichtchouk et al. (2018). The skill of the QBO forecasts decreases substantially after the first 2 months of the forecast. This is shown to be sensitive to the parametrization of the tropical non-orographic gravity wave drag in the model (Polichtchouk et al. (2018), Polichtchouk et al. (2017)). The amplitude of the QBO tends to weaken through the forecast. Even though some S2S models, including the ECMWF's Integrated Forecasting System (IFS, Vitart, 2014), are already able to reproduce the QBO's effect on the polar vortex, they are still underestimating the effect on surface weather (Garfinkel et al. 2018). ~~Even though many models are already able to reproduce the QBO, they are still underestimating its teleconnection to the surface weather (Scaife et al. 2014).~~ Furthermore, the anomalous QBO disruption in winter 2015/2016 was not forecasted by the models (Newman et al. 2016).

In this paper, we first verify the raw and the mean bias-corrected surface temperature reforecasts of the ECMWF's ERFs for forecast weeks 1 to 6 over Northern Europe against ERA-Interim surface temperature re-analysis (Dee et al. 2011). After that, our aim is to find out which stratospheric observations available at the start of the forecast are followed by a statistically significantly weaker AO index. For this, we explore the observed daily AO index during boreal winters 1981-2016, 1–2 weeks, 3–4 weeks, and 5–6 weeks after different phases of QBO and strengths of the observed stratospheric zonal mean winds. According to the observed daily AO index, after different phases of QBO, and strengths of the observed stratospheric zonal mean winds, we define a Stratospheric Wind Indicator (*SWI*), which is a novel indicator for the strength of the AO index in the following 1 to 6 weeks. For a statistically significantly weaker mean AO index, the *SWI* is defined as SWI_{neg}; otherwise,

SWI is defined as SWI_{plain}. Further, we study the mean surface temperature anomalies observed in Northern Europe 1–2 weeks, 3–4 weeks, and 5–6 weeks after SWI_{neg} versus SWI_{plain} and utilize these in post-processing the mean of the temperature forecasts of ECMWF reforecasts. Finally, we compare the *SWI* based post-processed ECMWF reforecasts with the mean bias-corrected ECMWF reforecasts. Our paper is constructed as follows: First, we present the datasets and methods. Then, we present results about the selection of the SWI_{neg} and SWI_{plain} and the skill scores of the forecasts without post-processing and with post-processing. In the Discussions and Conclusions section, we present our view on our findings and the possible next steps.

2 Datasets and Methods

We verified and post-processed ERFs of the ECMWF's ~~Integrated Forecasting System (~~IFS Cycle 43r1~~,~~ (Vitart, 2014), which belongs to the models of the Sub-seasonal to seasonal ~~scale~~ (S2S) prediction project of the World Weather Research Program/World Climate Research Program (Vitart et al. 2017). These forecasts are run twice a week, on Mondays and Thursdays, in a horizontal resolution of 0.4 degrees. We first studied the weekly mean temperatures of the Monday runs over Northern Europe (52° N to 71.2° N and 10° E to 33.2° E) with lead times of 1 to 6 weeks, here called ~~in~~ forecast weeks 1 to 6. We verified the 20 years × 52 weeks = 1040 reforecasts (11 members ensemble) for 1997–2016 run for the same dates as the operational forecasts, i.e., as Mondays in 2017. The weekly averages of the raw, mean bias-corrected (Section 2.2), and post-processed (Sections 2.3 and 2.4) surface temperature forecasts over Northern Europe were verified against ERA-Interim 1981–2016 temperature re-analyses (Dee et al. 2011). Years 1981–2010 of the ERA-Interim data were used as the climatological reference period and as the statistical/climatological forecast.

2.1 Skill scores of the forecasts

A commonly used measure for the probabilistic forecasts is the continuous ranked probability score (CRPS, Hersbach 2000) calculated by the following Eq. (1):

$$CRPS = \int |F(y) - F_o(y)|^2 dx, \quad (1)$$

where $F(y)$ and $F_o(y)$ are the cumulative distribution functions of the forecast and the observation, respectively.

The CRPSs were calculated by the R package 'ScoringRules' (Jordan et al. 2018) for the ECMWF's reforecast ($CRPS_{rf}$) and the climatological forecasts (ERA-Interim weekly mean temperatures in 1981–2010), which were used as the reference ($CRPS_{clim}$). As the ensemble size of the reforecasts, *m*, was only 11, and the ensemble size of the operational forecasts of the ECMWF's IFS, *M*, was 51, the expected CRPS, the $CRPS_{RF}$ of the ECMWF's reforecast ~~$CRPS_{rf}$~~ was calculated for 51 members using equation 26 in Ferro et al. (2008):

$$CRPS_{RF} = \frac{m(M+1)}{M(m+1)} CRPS_{rf} \quad (2).$$

We calculated the annual means of the expected $CRPS_{RF}$ across all weeks (1 to 52) of the years 1997-2016 reforecasts. These annual means were computed separately for lead times of 1 week, 2 weeks, 3 weeks, 4 weeks, 5 weeks, and 6 weeks, here called forecast week 1, forecast week2, forecast week 3, forecast week 4, forecast week 5, and forecast week 6, respectively. Further, the skill scores of the annual mean CPRSs, the annual mean CRPSSs, for each lead time were calculated as follows:

$$CRPSS = 1 - \frac{CRPS_{RF}}{CRPS_{clim}} \quad (3).$$

~~The annual mean CRPSSs were calculated for the ECMWF's reforecasts (1997–2016) for forecast weeks 1 to 6.~~ The statistical significances of each forecast week's annual mean CRPSS was determined for each grid point. The p-value with the null hypothesis that the CRPSS is zero was calculated by bootstrap resampling procedure with replacement and a sample size of 5000 for significance level 0.05.

2.2 Bias-correction of the ensemble mean

The mean bias-correction (as in Buizza and Leutbecher 2015, eq. 7a) removed the mean bias computed from the ensemble reforecasts for the 20 years (1997–2016) depending on the forecast week date. For the 1997–2016 reforecasts, the average bias was calculated considering 19 × 11 × 5 = 1045 ensemble reforecast members: 11 members' reforecast with initial dates defined by five weeks centred on the forecast week date for the 19 years reforecasts (1997–2016 excluding the reforecast year). The mean bias-corrected weekly mean temperatures were verified against the ERA-Interim data by calculating the annual mean CRPS separately for each lead time, i.e., forecast weeks ~~(~~1 to 6~~)~~. The skill scores of the mean bias-corrected forecasts and their statistical significance were calculated as explained in Section 2.1.

2.3 Definition of the stratospheric wind indicator (*SWI*)

As numerous observational and modelling studies have shown, the stratospheric polar vortex influences the weather in the Northern Hemisphere during boreal winter; strong polar vortex coincides more often with positive AO index and mild surface weather in Northern Europe, whereas weak polar vortex is more often followed by negative AO index and cold air outbreaks (Thompson and Wallace 1998, 2001, Kidston et al. 2015 and references therein). We aimed to find stratospheric precursors for a statistically significantly weaker AO index available at the start of the forecast. The ~~observed~~ daily surface AO indexes were downloaded from the National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center (CPC). We used two stratospheric wind data sets for the precursors of the AO index. The first data were the daily ZMZW at 60° N and 10 hPa during 1981–2016 of the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2; Rienecker et al. 2011) reanalysis data provided by the National Aeronautics and Space Administration (NASA). The second data were the monthly mean zonal wind components at levels 70 hPa, 50 hPa, 40 hPa, 30 hPa, 20 hPa, 15 hPa, and 10 hPa from the Singapore radio soundings, during 1981–2016, provided by the Free University of Berlin, representing the equatorial stratospheric monthly mean zonal wind components, the QBO (Naujokat 1986),

Formatted: Subscript

Formatted: Font: (Default) +Headings (Times New Roman)

~~We examined the observed daily AO indexes during the 1–2 weeks, 3–4 weeks, and 5–6 weeks following different phases of QBO and strengths of the stratospheric winds.~~As Scaife et al. (2014) demonstrated a more negative AO in the easterly QBO at 30 hPa compared to the westerly QBO at 30 hPa, ~~Holton and Tan (1980, 1982) demonstrated that the geopotential height at high latitudes was significantly lower during westerly QBO compared to the easterly QBO. Therefore,~~ we explored the ~~daily~~ AO index 1–6 weeks after following predictors:

- westerly QBO at 30 hPa, the *WQBO*, ~~and~~
- easterly QBO at 30 hPa, the *EQBO*, ~~using the QBO winds at 30 hPa.~~
- *EQBO* with~~We also examined the effect of~~ the maximum of the monthly mean zonal wind components of the QBO between 70 hPa and 10hPa restricted to 7ms^{-1}, 10ms^{-1}, and 13ms^{-1}, ~~during *EQBO*. Moreover, we explored the daily AO index after~~
- the daily ZMZW at 60° N and 10 hPa during the last 10 days of the previous month falling below its overall wintertime (November–March 1981–2016) 10th percentile, corresponding a value of 3.8m/s, indicating a weak polar vortex already at the start of the forecast.

The statistical significance of the difference between the AO index following two different stratospheric situations, e.g., the *EQBO* and the *WQBO*, was determined using a two-sided Student's t-test with the null hypothesis that there is no difference. ~~We used the~~The most statistically significant predictors for weaker AO indexes observed 1–2 weeks, 3–4 weeks, and 5–6 weeks after these stratospheric situations, were used to define a *SWI* to be SWI_{neg}; otherwise, it was defined as SWI_{plain} for the beginning of each winter month (November–February) in 1981–2016.

2.4 Utilizing the stratospheric winds indicator (*SWI*) in forecasting

In this section, we investigated the observed and reforecasted surface temperature anomalies 1–2 weeks, 3–4 weeks, and 5–6 weeks after SWI_{neg} and SWI_{plain} defined in Section 2.3. First, we calculated the observed two-week mean temperature anomalies of the ERA-Interim reanalyses (Dee et al. 2011) of the 1–2 weeks, the 3–4 weeks, and the 5–6 weeks from the beginning of the January, February, November, and December in 1981–2016 in Northern Europe. Subsequently, we divided the observed two-week mean temperature anomalies to sets of anomalies, representing SWI_{neg} and SWI_{plain} according to the previous weeks' ~~month's~~ stratospheric wind conditions~~observations~~. Thereafter, we determined the statistical significance of the difference between the surface temperatures after SWI_{neg} and SWI_{plain} using a two-sided Student's t-test with the null hypothesis that there is no difference between SWI_{neg} and SWI_{plain}. This same procedure to define the difference between the surface temperatures after SWI_{neg} and SWI_{plain} was used for the ERA-Interim reanalyses for the period 1997–2016 to see how the selection of a shorter period affects the temperature anomalies. Further, the mean surface temperature anomalies 1–2 weeks, 3–4 weeks, and 5–6 weeks after SWI_{neg} and SWI_{plain} in the ECMWF reforecasts run of the first week of November–February 1997–2016 were defined to examine how the model reproduced the anomalies.

Formatted: List Paragraph, Bulleted + Level: 1 + Aligned at: 0.63 cm + Indent at: 1.27 cm

Formatted: Font: Italic

Formatted: Superscript

Formatted: Superscript

Formatted: Superscript

For post-processing the ECMWF reforecasts, we calculated TA_{SWIneg} and $TA_{SWIplain}$ representing mean temperature anomalies in November–February 1981-2016 after SWI_{neg} and SWI_{plain}, respectively. ~~In order to decrease the effect of the time period used for defining mean anomalies, the TA_{SWIneg} and $TA_{SWIplain}$ were calculated from the ERA-Interim 1981–2016 two-week mean temperature anomalies by dividing thirty five years' data (1981–2016 excluding the reforecast year) to 5-year-long periods. One 5-year-long period was left out, and the mean anomalies representing SWI_{neg} and SWI_{plain} were calculated of the remaining 30 years. This was repeated seven times by changing the 5-year-long period left out.~~ The means, TA_{SWIneg} and $TA_{SWIplain}$ ~~of the achieved seven mean temperature anomalies~~ representing SWI_{neg} and SWI_{plain} were calculated separately for each ~~grid~~ 0.4° × 0.4° grid point over Northern Europe.

For the post-processing of the ECMWF reforecasts, we first defined the *SWI* either SWI_{neg} or SWI_{plain} at the start of the forecast according to previous weeks' ~~months'~~ stratospheric wind conditions ~~observations~~. According to the *SWI*, we added either TA_{SWIneg} or $TA_{SWIplain}$ to the ERA-Interim mean temperature during 1981–2016, corresponding to forecast weeks 1–2, 3–4, and 5–6 to get a SWI_{neg} and SWI_{plain} based mean temperatures, T_{SWIneg} and $T_{SWIplain}$, for weeks 1–2, 3–4, and 5–6, respectively. The T_{SWIneg} and $T_{SWIplain}$ were used in post-processing the ECMWF reforecasts' mean bias-corrected ensemble members, T_{BC}, by calculating a weighted average, T_{SWI_BC}, for SWI_{neg} as follows:

$$T_{SWI\_BC} = (1 - k_{SWI}) * T_{BC} + k_{SWI} * T_{SWIneg} \quad (4)$$

And for SWI_{plain},

$$T_{SWI\_BC} = (1 - k_{SWI}) * T_{BC} + k_{SWI} * T_{SWIplain} \quad (5)$$

where T_{SWI_BC} was a post-processed ensemble member. k_{SWI} was the weight of the T_{SWIneg} or $T_{SWIplain}$, which was tested between 0–1 and defined according to the best improvement in the skill scores of the post-processed forecast. By Eq. (4) and Eq. (5), we adjusted each ensemble member with the same weight, and hence, the original spread of the ECMWF reforecasts remained unchanged. The skill scores of the *SWI* based post-processed forecasts, and their statistical significance, were calculated as explained in Section 2.1.

3 Results

3.1 Skill scores of the forecasts

The annual mean of the expected CRPSS and its 95% level of confidence of the raw and the mean bias-corrected (Section 2.2) weekly mean temperature of the ECMWF reforecasts for 1997–2016 are displayed in Figure 1. In grid points where the CRPSS was higher than zero and the confidence level was higher than 95% (dotted areas), the reforecasts were statistically significantly better than just the statistical forecast based on 1981–2010 climatology. Figure 1 illustrates that for forecast weeks 1–6 the mean bias-corrected ERF reforecasts were on average significantly better than climatology. The annual mean CRPSS values

show that in forecast weeks 1–3 the CRPSSs are for the most part above 0.1, whereas on in forecast weeks 4–6 they are mostly lower, between 0 and 0.1.

3.2 The stratospheric observations and the thereafter observed AO index and surface temperature

Figure 2 shows boxplots of the observed mean~~minimum~~ of the daily AO index 1–2 weeks, 3–4 weeks, and 5–6 weeks after different phases of QBO and restrictions in the strength of the stratospheric winds in 1981–2016. The first box (brown) represents the mean ~~minimum~~ AO indexes after all the cases in 1981–2016 November–February, i.e., 36 years * 17 weeks=612 cases.~~, i.e., 36 years * 4 months =144 cases.~~ The second and third boxes show the mean ~~minimum~~ AO indexes after easterly (*EQBO*, blue) and westerly (*WQBO*, pink) QBO at the 30 hPa level, respectively. The p-value written below each boxplot pair indicates the likelihood of such a pair of distributions arising from a random sampling of a single distribution as given by a Student's t-test, i.e., p-values less than 0.05 indicate that the means of the data sets differ significantly at the 95% level of confidence. The median and the mean of the mean ~~minimum~~ AO indexes 1–2 weeks, 3–4 weeks, and 5–6 weeks after *EQBO* were statistically significantly lower than after *WQBO*. ~~However, this was statistically significant at the 95% confidence level only in weeks 1–2.~~ The *EQBO* (blue) box shows all the cases of *EQBO* with no restriction in the QBO's monthly mean zonal wind components, whereas the fourth, the sixth, and the eighth blueish boxes show the mean ~~minimum~~ AO indexes after *EQBO* with all the maximum QBO's monthly mean zonal wind components between levels 70…10hPa being below 13 m/s, 10 m/s, and 7 m/s, respectively. Restricting the *EQBO* cases by a maximum of the QBO's monthly mean zonal wind components in levels 70…10hPa decreased the median of the mean ~~minimum~~ AO during the following 1–2, 3–4, and 5–6 weeks.~~; however, only for the QBO's monthly mean zonal wind components less than 10 m/s, the minimum AO index was still at the 95% confidence level statistically significantly lower than in the rest of the data.~~

The 10th box (yellow) in Fig. 2 shows the mean ~~minimum~~ AO indexes after cases where the daily ZMZW at 60° N and 10 hPa was below its 10th percentile (3~~4~~.8m/s) during the last 10 ~~last~~ days preceding the start of the forecast, ~~of the previous month,~~ corresponding to cases with weak polar vortex already at the start of the forecast. The observed mean ~~minimum~~ AO index was statistically significantly weaker at the 9~~5~~% confidence level ~~weaker~~ 1–2 weeks, ~~and~~ 3–4 weeks, and 5–6 weeks after the daily ZMZW at 60° N and 10 hPa had ~~, which was~~ been below its~~their~~ overall wintertime 10th percentile (indicating a weak polar vortex). ~~In weeks 5–6, the AO index was also weaker, but it was statistically insignificant at the 95% level.~~

Aiming to select stratospheric precursors indicating weak AO with the greatest statistical significance, we defined the *SWI* to be negative in cases when the QBO was easterly at 30 hPa and the QBO's monthly mean zonal wind components in levels 70…10hPa were weaker than 10m/s and/or if the daily ZMZW at 60° N and 10 hPa during the 10 last days of the previous month fell below its overall wintertime 10th percentile. In other cases, the *SWI* was defined as plain. This decision tree for the *SWI* is depicted in Fig. 3. The means of the mean ~~minimum~~ AO index after SWI_{neg} and SWI_{plain} (in 1981–2016) were statistically

significantly different using a Student's t-test, with lower AO index more common 1–2 weeks, 3–4 weeks, and 5–6 weeks after SWI_{neg} than after SWI_{plain} (see Fig. 2 for the p-values).

Figure 4 shows the observed (periods 1981–2016 and 1997–2016) and model forecasted (the period 1997–2016) mean temperature anomalies of the weeks' 1–2, 3–4, and 5–6 in November–February after SWI_{neg} and SWI_{plain}. The observations showed on average lower mean temperatures for the weeks' 1–2, 3–4 and 5–6 after SWI_{neg} (Fig. 4ab-c and 4gh-i). The reforecasts also showed cold anomalies after SWI_{neg} (Fig. 4mn-o) but for forecast weeks 3–4 and 5–6 weaker than the observed ones. Further, the observations showed on average higher mean temperatures for weeks 1–2, 3–4, and 5–6 after SWI_{plain} (Fig. 4d-f and 4j-l). This warm anomaly was well forecasted only to some degree in the forecasts weeks 1–2 (Fig. 4p), and it was reasonable well forecasted totally absent in forecast weeks 3–4 (Fig. 4q) and 5–6 (Fig. 4r). The mean temperature anomalies 31-6 weeks after SWI_{neg} (Fig. 4b4a-c) and SWI_{plain} (4e4d-f) during 1981–2016 were statistically significantly different using a Student's t-test, with anomalously cold surface temperatures more common 31-6 weeks after SWI_{neg}. When examining the years 1997–2016 (Fig. 4h4g-i and Fig. 4k4j-l), which was the reforecast period, the temperature anomalies were of the same sign than during the longer 1981–2016 period (Fig. 4b4a-c and Fig. 4e4d-f). but weaker and not statistically significant all over the Northern Europe.

3.3 The *SWI* and the forecasted mean temperatures

The mean temperature anomalies in Fig. 4(a-f) for Northern Europe were used for the *SWI* based post-processing as described in Section 2.4. The CRPSS of the mean temperature of the forecast weeks 1–2 were not improved by the *SWI* (no figure), whereas the CRPSSs of the mean temperatures of the forecast weeks 3–4 and 5–6 were improved by the *SWI* based post-processing (Fig. 5a and 5b). The best median CRPSS was achieved by k_{SWI}=0.45, for both forecast weeks 3–4 and by k_{SWI}=0.6 for forecast weeks 5–6. Figure 6 shows the forecasts skill of the mean temperature of the forecast weeks 3–4 and weeks 5–6 forecasted by the mean bias-corrected reforecasts alone (Fig. 6a-b) and by the mean bias-corrected reforecasts together with the *SWI* based post-processed forecast (k_{SWI} =0.5, Fig. 6c-d). By using the *SWI* based post-processing to the ECMWF forecasts, the CRPSSs for weeks 3–4 and weeks 5–6 were slightly improved and the area of these forecasts being significantly better than just the climatological forecast was expanded. The forecast skills for the weeks 3–4 and 5–6 post-processed by Eq. (4) and Eq. (5) were not sensitive to the period (within 1981–2016) used for defining SWI_{neg} and SWI_{plain} temperature anomalies. For instance, we tested periods 1981–2000 and 2001–2016 and got almost the same CRPSSs for Northern Europe (no figure).

4 Discussion and Conclusions

Based on ECMWF's extended-range reforecasts for the period 1997–2016, we found that the weekly mean surface temperature forecasts over Northern Europe were on average significantly better than just the climatological forecast in weeks 1–6, however, in weeks 4–6, the CRPSSs were quite low, mostly between 0 and 0.1.

We showed that in addition to the previously demonstrated more negative AO ~~weaker polar vortex~~ during easterly QBO in comparison to westerly QBO at 30 hPa (Scaife et al. 2014)~~e.g., Holton and Tan 1980, Garfinkel et al. 2018)~~, the mean AO index was sensitive to the maximum strength of the QBO's monthly mean zonal wind components in levels 70…10hPa during the easterly QBO at 30 hPa~~ affected the observed AO index 1–6 weeks later~~. Based on observations, we found that the mean ~~minimum~~ AO index was statistically significantly weaker 1–2 weeks, 3–4 weeks, and 5–6 weeks after the monthly mean QBO was easterly at 30 hPa, and all the QBO's monthly mean zonal wind components in levels 70…10hPa were less than 10 m/s. We also found that the mean ~~minimum~~ AO index was statistically significantly weaker 1–2 weeks~~, and~~ 3–4 weeks, and 5–6 weeks after the daily ZMZWs at 60° N~~,~~ and 10 hPa had been below their overall wintertime 10th percentile (indicating a weak polar vortex). ~~In weeks 5–6, the AO index was weaker but statistically insignificant.~~

Selecting the SWI_{neg} to include the both above-mentioned situations, ~~the level of statistical significance of a weaker AO index during the next 1–2, 3–4, and 5–6 weeks decreased in comparison to using only one of these situations. Our definition of SWI_{neg}~~ resulted in a statistically significantly weaker AO index within the following 1-6 weeks in comparison to the rest of the data, defined as SWI_{plain}. As negative AO index enables cold air outbreaks to Northern Europe (Thompson et al. 2002, Tomassini et al. 2012) and positive AO index tends to bring milder and wetter than average weather to Northern Europe (Limpasuvan et al. 2005), we tested the predictor SWIneg/SWIplain as a predictor of mean surface temperature in Northern Europe for forecast weeks 3–6. ~~Also,~~ We found that the mean surface temperature anomalies in Northern Europe in November–February in 1981–2016 after SWI_{neg} and SWI_{plain} were statistically significantly different, with anomalously cold surface temperatures more common 3–6 weeks after SWI_{neg}. The mean temperature anomalies corresponding to SWI_{neg}/SWI_{plain} at the start of the forecast were used in post processing the ECMWF's mean temperature reforecast for weeks 3–4 and 5–6 in Northern Europe during boreal winter, and thereby, those weeks' forecast skills were slightly improved.

This study demonstrates that the QBO-polar vortex connection should be better integrated into the extended-range surface temperature forecasts over Northern Europe. The *SWI* based post-processing method introduced in this paper could also be tested for other northern areas affected by the polar vortex and to precipitation and windiness forecasts, and it could be further developed by, e.g., the Madden-Julian-Oscillation (Madden and Julian 1994; Zhang 2005; Jiang et al. 2017; Vitart 2017; Vitart and Molteni 2010; Robertson et al. 2018, Cassou 2008). In this study, the effect of global warming was not filtered from the temperature anomalies used for statistical post-processing. In future work, the impact of filtering the effect of global warming could be tested. ~~In our studies, the use of monthly mean stratospheric observations restricted the studies to post processing only the forecasts made on the first week of each month. The~~Moreover, the next step would be looking for the stratospheric signal from the forecast model~~, which would also make it possible to post-process forecasts made every week~~.

Data availability. ERA-Interim data available at https://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/ (last accessed 24 June 2019). ECMWF reforecasts data available at https://apps.ecmwf.int/mars-catalogue/ (last accessed 28 June 2019). AO indexes data available at https://www.cpc.ncep.noaa.gov/products/precip/CWlink/daily_ao_index/ao.shtml (last accessed 24 June 2019). The daily ZMZW at 60° N and 10 hPa data available at https://acd-ext.gsfc.nasa.gov/Data_services/met/ann_data.html (last accessed 24 June 2019). The QBO data data available at https://www.geo.fu-berlin.de/met/ag/strat/produkte/qbo/qbo.dat (last accessed 24 June 2019). The data of Figures 1–2 and 4–6 available at https://etsin.fairdata.fi/dataset/6417610e-5f1c-45fd-9814-02209ba38051 ~~https://github.com/fmidev/sixweeks~~.

Competing interests. The authors declare that they have no conflict of interest.

Author contributions. NK designed the study, analysed the results and prepared the manuscript with contributions from all co-authors. OH participated in the study design and analysing the results. MK contributed to the discussions and fine-tuned the experiments. DSR contributed to the discussions and to the interpretation of the results. HJ provided supervision during the experiments and writing. HG contributed to the study design and was in charge of the management and the acquisition of the financial support for the CLIPS-project leading to this publication.

Acknowledgements. We wish to thank Academy of Finland for funding the project (number 303951 SA CLIPS). We also acknowledge the ECMWF for monthly forecast data and ERA-Interim data, NOAA/CPC for providing the AO index data, NASA for providing 10hPa wind data, and Free University of Berlin for providing the QBO data. We thank the CLIPS team and developers of the R cran calculation package 'ScoringRules'. We thank the three anonymous reviewers for their good and constructive comments.

References

Baldwin, M. P., and Dunkerton, T.J.: Propagation of the Arctic Oscillation from the stratosphere to the troposphere, J. Geophys. Res, 104, D24, 30937-30946, https://doi.org/10.1029/1999JD900445, 1999.

Baldwin, M. P., and Dunkerton, T. J.: Stratospheric harbingers of anomalous weather regimes. Science, 294, 581–584, doi:10.1126/science.1063315, 2001.

Baldwin, M. P., Gray, L. J., Dunkerton, T. J., Hamilton, K., Haynes, P. H., Randel, W. J., et al.: The quasi-biennial oscillation, Rev. Geophys., 39(2), 179–229, 2001.

Buizza, R. and Leutbecher, M.: The forecast skill horizon, Q. J. R. Meteorol. Soc., 141, 3366–3382, doi:10.1002/qj.2619, 2015.

Butler, A. H., Seidel, D. J., Hardiman, S. C., Butchart, N., Birner, T., and Match, A.: Defining sudden stratospheric warmings, Bull. American Meteor. Soc., 96, 1913-1928, doi: http://dx.doi.org/10.1175/BAMS-D-13-00173.1, 2015.

Cassou C.: Intraseasonal interaction between the Madden–Julian Oscillation and the North Atlantic Oscillation. Nature, 455, 523–527, 2008.

Chambers, J. M., Cleveland, W. S., Kleiner, B, and Tukey, P.A.: Graphical Methods for Data Analysis, The Wadsworth statistics/probability series. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1983.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. Roy. Meteorol. Soc., 137, 553–597, 2011.

Ervasti, T., Gregow, H., Vajda, A., Laurila, T. K., and Mäkelä, A.: Mapping users' expectations regarding extended-range forecasts. Adv. Sci. Res., 15, 99–106, doi: 10.5194/asr-15-99-2018, 2018.

Ferro C. A. T., Richardson, D. S., and Weigel, A. P.: On the effect of ensemble size on the discrete and continuous ranked probability scores. Meteorol. Appl. 15: 19–24, doi: 10.1002/met.45, 2008.

Garfinkel, C. I., Schwartz, C., Domeisen, D. I. P., Son, S-W., Butler, A. H., and White, I. P.: Extratropical stratospheric predictability from the Quasi-Biennial Oscillation in Subseasonal forecast models, J. Geophys. Res: Atmospheres, doi: 10.1029/2018JD028724, 2018.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems. Wea. Forecasting, 15, 559–570, doi:10.1175/1520-0434(2000)015,0559: DOTCRP.2.0.CO;2, 2000.

Gray, L. J., Anstey, J. A., Kawatani, Y., Lu, H., Osprey, S., and Schenzinger, V.: Surface impacts of the Quasi Biennial Oscillation, Atmos. Chem. Phys., 18, 8227–8247, https://doi.org/10.5194/acp-18-8227-2018, 2018.

Holton, J. R. and Tan, H. C.: The influence of the equatorial quasi-biennial oscillation on the global circulation at 50mb, J. Atmos. Sci., 37, 2200–2208, 1980.

Holton, J. R. and Tan, H. C.: The quasi-biennial oscillation in the Northern Hemisphere lower stratosphere, J. Meteor. Soc. Japan, 60, 140–148, 1982.

Kidston, J.; Scaife, A. A.; Hardiman, S. C.; Mitchell, D. M.; Butchart, N.; Baldwin, M. P., and Gray, L. J.: Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. Nature Geoscience, 8(6), 433–440, 2015.

Limpasuvan, V., Hartmann, D. L., Thompson, D. W. J., Jeev, K., and Yung, Y. L.: Stratosphere-troposphere evolution during polar vortex intensification. J. Geophys. Res., 110, D24101, doi: 10.1029/2005JD006302, 2005.

Madden, R. A., and Julian, P. R.: Observations of the 40–50-day tropical oscillation—A review. Mon. Wea. Rev., 122, 814–837, 1994.

Jiang, Z., Feldstein, S. B., and Lee S.: The relationship between the Madden–Julian Oscillation and the North Atlantic Oscillation. Q. J. R. Meteorol.Soc. 143: 240–250, January 2017 A DOI:10.1002/qj.2917, 2017.

Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C., and Liniger, M. A.: Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. J. Geophys. Res: Atmospheres, 123, 7999–8016. 2018.

Jordan A., Krueger F., and Lerch, S.: Evaluating Probabilistic Forecasts with scoringRules. Journal of Statistical Software, forthcoming, 2018.

Naujokat, B.: An update of the observed quasi-biennial oscillation of the stratospheric winds over the tropics. J. Atmos. Sci., 43, 1873–1877. 1986.

Newman, P. A., L. Coy, S. Pawson, and L. R. Lait: The anomalous change in the QBO in 2015–2016, Geophys. Res. Lett., 43, 8791–8797, 2016.

Polichtchouk, I., et al., 2017: What influences the middle atmosphere circulation in the IFS? ECMWF Technical Memorandum No. 809.

Polichtchouk, I., Shepherd, T. G., Byrne, N. J.: Impact of Parametrized Nonorographic Gravity Wave Drag on Stratosphere-Troposphere Coupling in the Northern and Southern Hemispheres, Geophys. Res. Lett., 45, 8612-8618, doi: 10.1029/2018gl078981, 2018.

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Emily Liu, J. B., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G. K., Bloom, S., Chen, J., Collins, D., Conaty, A., da Silva, A., Gu, W., Joiner, J., Koster, R. D., Lucchesi, R., Molod, A., Owens, T., Pawson, S., Pegion, P., Redder, C. R., Reichle, R., Robertson, F. R., Ruddick, A. G., Sienkiewicz, M., Woollen, J.: MERRA: NASA's modern-era retrospective analysis for research and applications. J. Clim. 24: 3624–3648. https://doi.org/10.1175/JCLI-D-11-00015.1, 2011.

Robertson, A. W., Camargo, S. J., Sobel, A., Vitart, F., and Wang, S.: Summary of workshop on sub-seasonal to seasonal predictability of extreme weather and climate. npj Climate and Atmospheric Science, 1, 8, doi: 10.1038/s41612-017-0009-1, 2018.

Scaife, A. A., et al.: Predictability of the quasi-biennial oscillation and its northern winter teleconnection on seasonal to decadal timescales, Geophys. Res. Lett., 41, 1752–1758, doi:10.1002/2013GL059160, 2014.

Shepherd T. G., Polichtchouk, I., Hogan, R., Simmons, A. J.: Report on Stratosphere Task Force, ECMWF Technical Memorandum n. 824, doi: 10.21957/0vkp0t1xx, 2018.

Thompson, D. W. J. and Wallace, J. M.: The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. Geophys. Res. Lett., 25, 1297– 1301, 1998.

Schoeberl, M. R.: Stratospheric warmings: Observations and theory. Rev. Geophys., 16, 521–538, 1978.

Thompson, D. W. J., Baldwin, M. P. and Wallace J. M.: Stratospheric connection to Northern Hemisphere wintertime weather: implications for prediction. J. Clim. 15, 1421–1428, 2002.

Thompson, D. W. J., and Wallace, J. M.: Regional Climate Impacts of the Northern Hemisphere Annular Mode. Science, 293, 85–89, 2001.

Tomassini, L., Gerber, E. P., Baldwin, M. P., Bunzel, F. and Giorgetta, M.: The role of stratosphere troposphere coupling in the occurrence of extreme winter cold spells over northern Europe. J. Adv. Model. Earth Syst., 4, M00A03, 2012.

Vitart F., and Molteni F.: Simulation of the MJO and its teleconnections in the ECMWF forecast system. Q. J. R. Meteorol. Soc., 136, 842–855, 2010.

Vitart F.: Evolution of ECMWF sub-seasonal forecast skill scores. Q. J. R. Meteorol. Soc., 140, 1889–1899, doi: 10.1002/qj.2256, 2014.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D., Xiao, H., Zaripov, R., and Zhang L.: The Subseasonal to Seasonal (S2S) Prediction Project Database. Bull. Amer. Meteor. Soc., 98, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1, 2017.

Vitart, F.: Madden-Julian Oscillation prediction and teleconnections in the S2S database: MJO prediction and teleconnections in the S2S database. Q. J. R. Meteor. Soc., 143, 2210–2220, 2017.

Watson, P. A., and L. J. Gray: How Does the Quasi-Biennial Oscillation Affect the Stratospheric Polar Vortex? J. Atmos. Sci.,71, 391–409, doi: 10.1175/JAS-D-13-096.1, 2014.

Zhang, C.: Madden-Julian Oscillation. Rev. Geophys., 43, RG2003, doi:10.1029/2004RG000158, 2005.

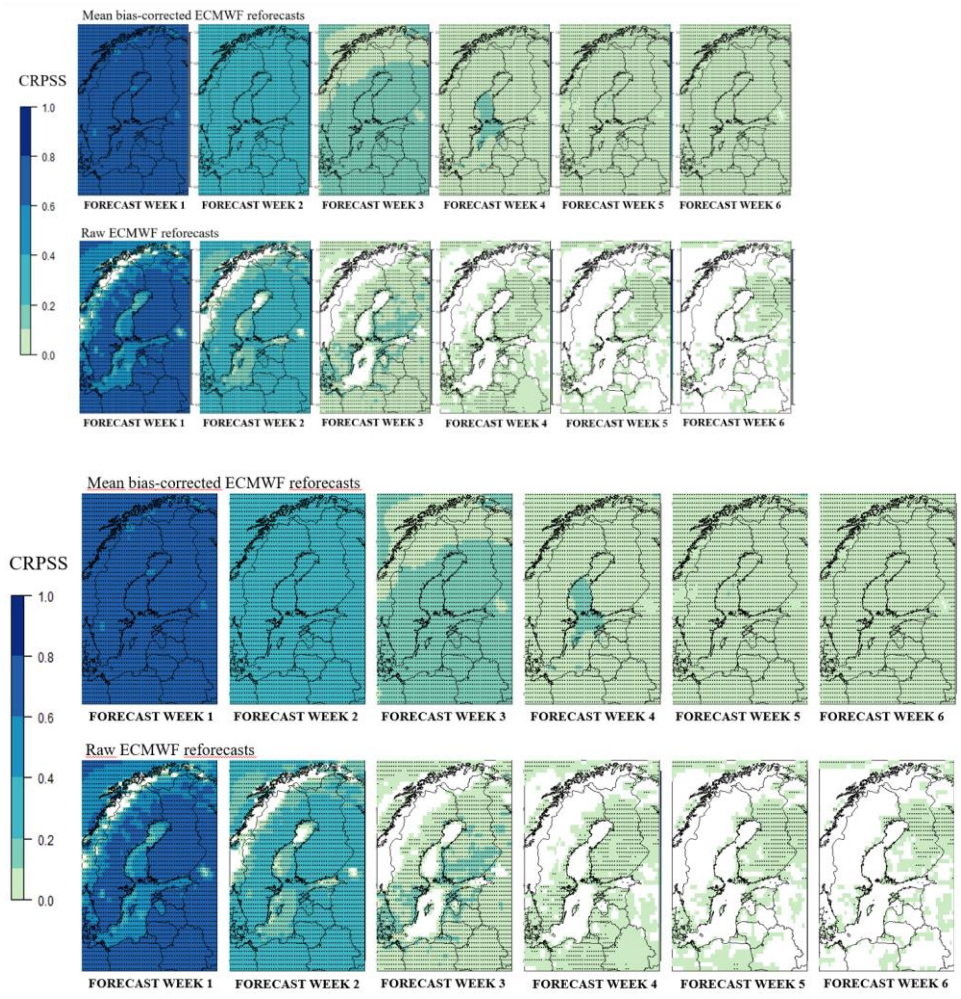
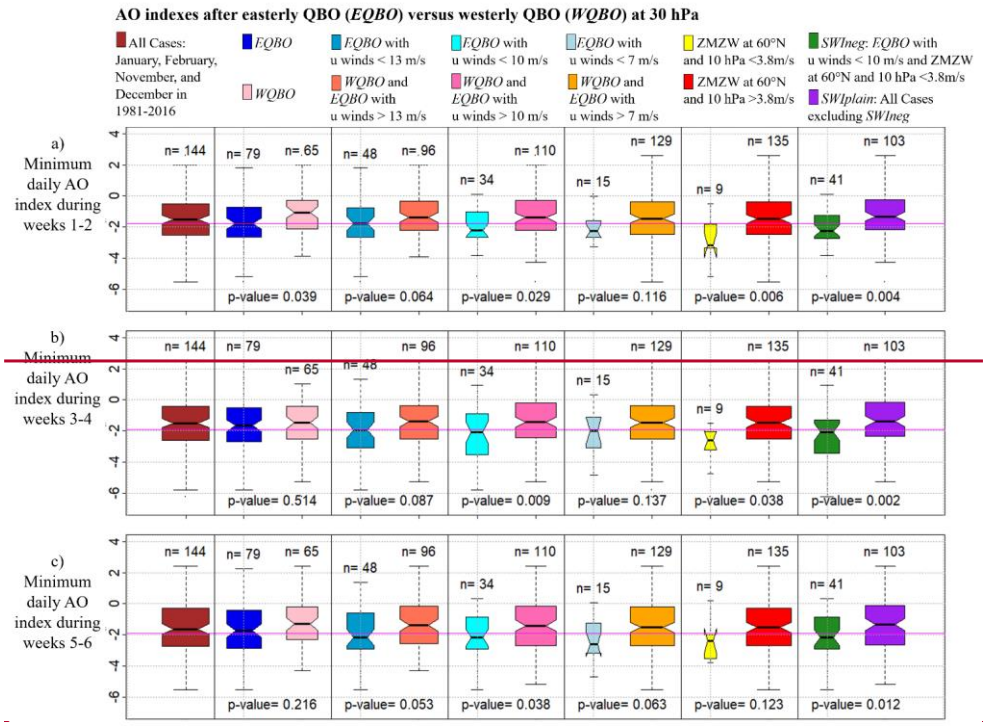


Figure 1: Annual mean of the expected CRPSS of the weekly mean temperature of the mean bias-corrected (upper row) and raw (lower row) ECMWF reforecasts for years 1997–2016 using ERA-Interim climatology of 1981–2010 as the reference. The dotted areas represent the 95% level of confidence that the CRPSS is above zero.


AO indexes after easterly QBO (EQBO) versus westerly QBO (WQBO) at 30 hPa
All Cases: January, February, November, and December in 1981-2016
EQBO
WQBO
EQBO with u winds < 13 m/s
WQBO and EQBO with u winds > 13 m/s
EQBO with u winds < 10 m/s
WQBO and EQBO with u winds > 10 m/s
EQBO with u winds < 7 m/s
WQBO and EQBO with u winds > 7 m/s
ZMZW at 60°N and 10 hPa <3.8m/s
ZMZW at 60°N and 10 hPa >3.8m/s
SWIneg: EQBO with u winds < 10 m/s and ZMZW at 60°N and 10 hPa <3.8m/s
SWIplain: All Cases excluding SWIneg
a) Minimum daily AO index during weeks 1-2
n= 144 n= 79 n=.65 n= 48 n= 96 n= 34 n= 110 n= 15 n= 129 n= 9 n= 135 n= 41 n= 103
p-value= 0.039 p-value= 0.064 p-value= 0.029 p-value= 0.116 p-value= 0.006 p-value= 0.004
b) Minimum daily AO index during weeks 3-4
n= 144 n= 79 n= 65 n= 48 n= 96 n= 34 n= 110 n= 15 n= 129 n= 9 n= 135 n= 41 n= 103
p-value= 0.514 p-value= 0.087 p-value= 0.009 p-value= 0.137 p-value= 0.038 p-value= 0.002
c) Minimum daily AO index during weeks 5-6
n= 144 n= 79 n= 65 n= 48 n= 96 n= 34 n= 110 n= 15 n= 129 n= 9 n= 135 n= 41 n= 103
p-value= 0.216 p-value= 0.053 p-value= 0.038 p-value= 0.063 p-value= 0.123 p-value= 0.012

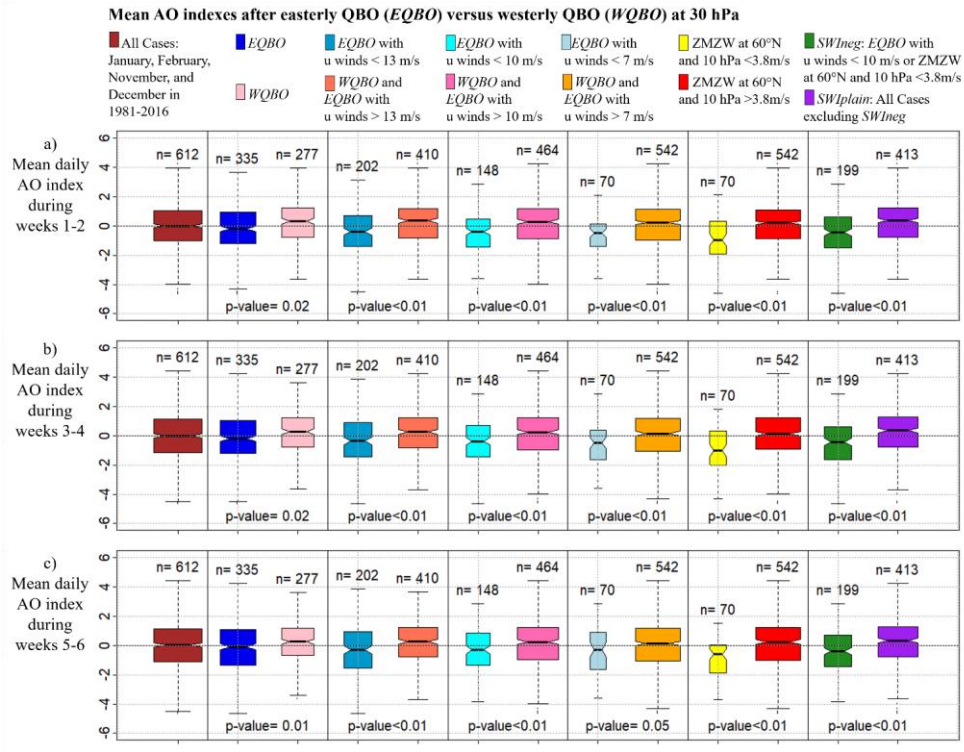


Figure 2: Observed ~~mean~~minimum daily AO index a) 1–2, b) 3–4 and c) 5–6 weeks after different stratospheric situations. The horizontal line dividing each box into two parts shows the median of the data, the ends of the box show the lower and upper quartiles, and the whiskers represent the highest and the lowest values excluding outliers. The *n* written above each box indicates the number of observations in each group. The widths of the boxes have been drawn proportional to the square-roots of *n*. The p-value written below each boxplot pair indicates the likelihood of such a pair of distributions arising from a random sampling of a single distribution as given by a Student's t-test, i.e., p-values less than 0.05 indicate that the means of the data sets differ significantly at the 95% level of confidence. The notches of each side of the boxes were calculated by R boxplot.stats. If the notches of two plots do not overlap, this is 'strong evidence' that the two medians differ (Chambers et al., 1983, p. 62). ~~The magenta line represents the mean minus one standard deviation of the daily AO index of all the cases.~~ ZMZW=zonal mean zonal wind. *SWI*=Stratospheric Wind Index.

Formatted: Font: Italic

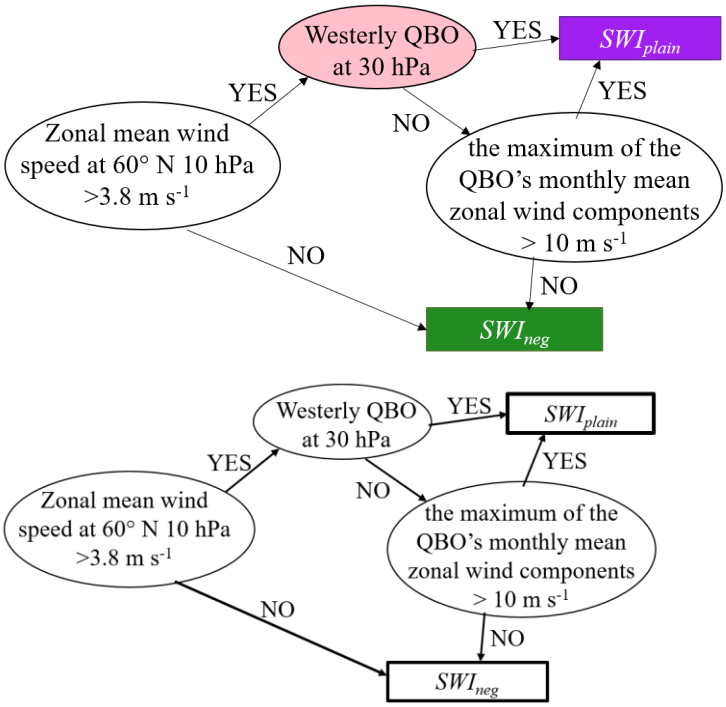
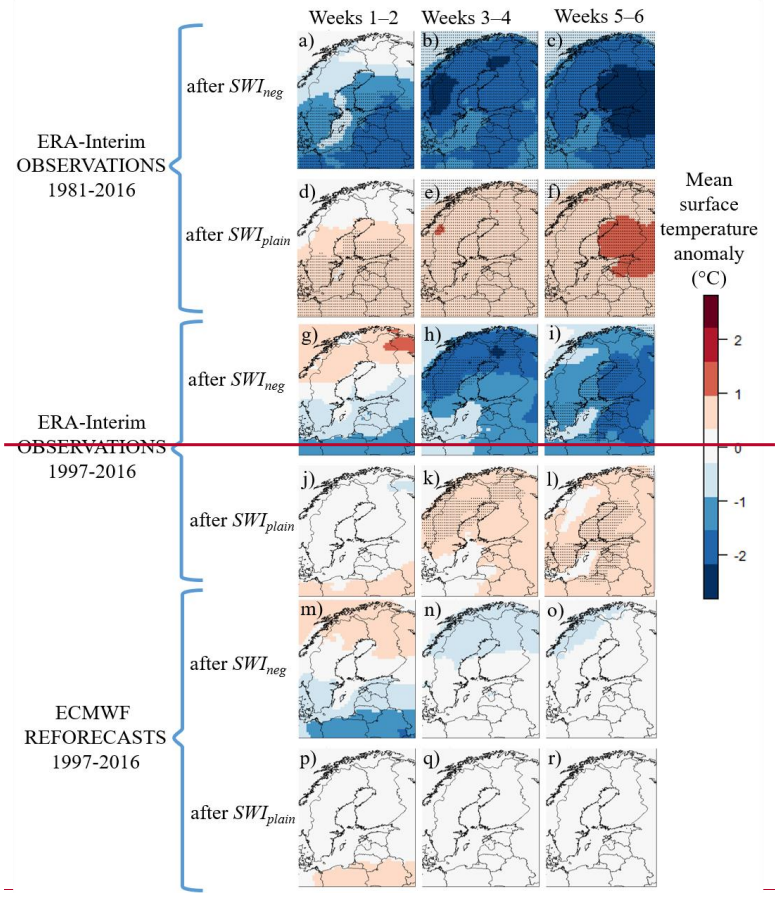


Figure 3: Decision tree of SWI_{neg}/SWI_{plain}.


Weeks 1–2
Weeks 3–4
Weeks 5–6
a)
b)
c)
d)
e)
f)
g)
h)
i)
j)
k)
l)
m)
n)
o)
p)
q)
r)
after SWI_{neg}
after SWI_{plain}
after SWI_{neg}
after SWI_{plain}
after SWI_{neg}
after SWI_{plain}
ERA-Interim OBSERVATIONS 1981-2016
ERA-Interim OBSERVATIONS 1997-2016
ECMWF REFORECASTS 1997-2016
Mean surface temperature anomaly (°C)
2
1
0
-1
-2

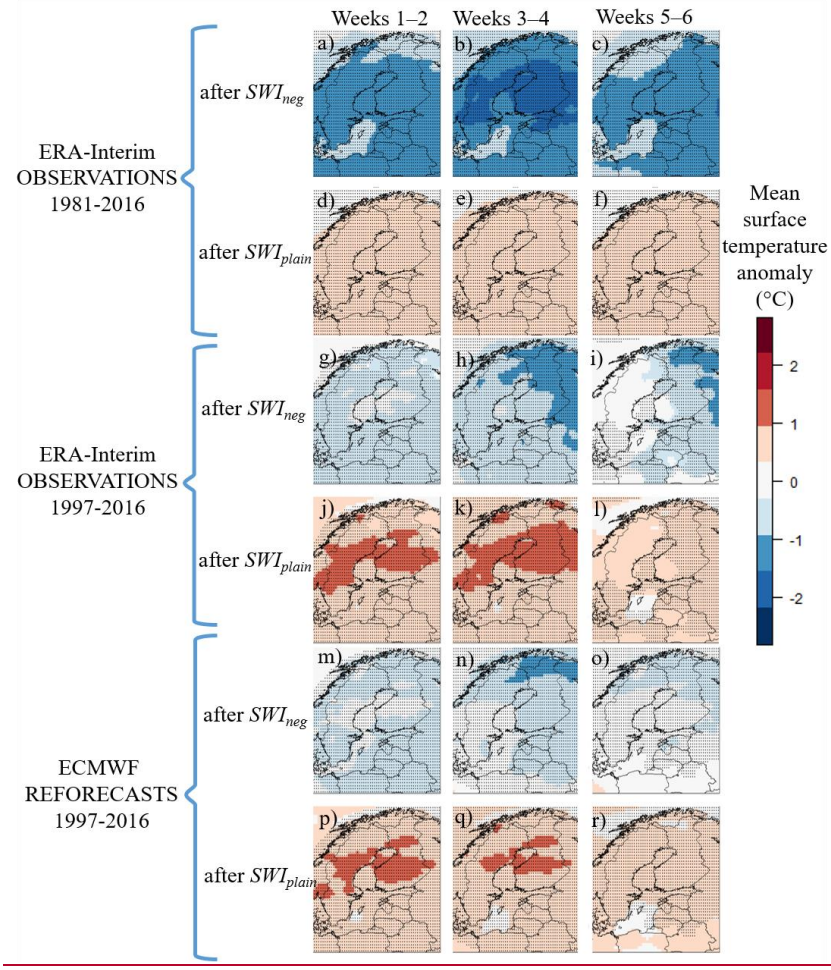
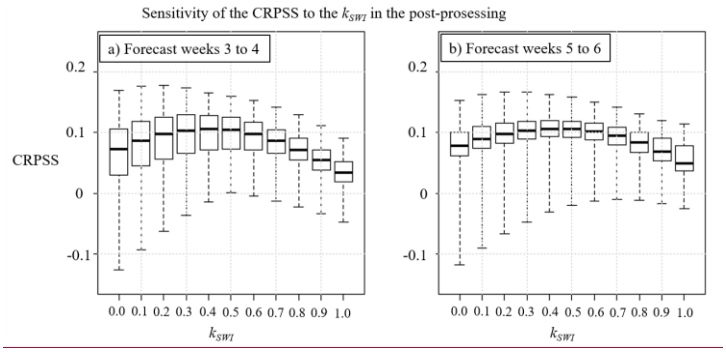


Figure 4. ERA-Interim observed (a-l) and ECMWF reforecasted (m-r) mean temperature anomalies in comparison to the 1981-2016 mean during boreal winters (November-February) in cases the previous month's *SWI* was negative (SWI_{neg}, covering about ~~28~~30% of the winter months) or plain (SWI_{plain}, covering about ~~72~~70% of the winter months). The dotted areas represent the 95% level of confidence where the means of surface temperature anomalies after SWI_{neg} and SWI_{plain} differ significantly.



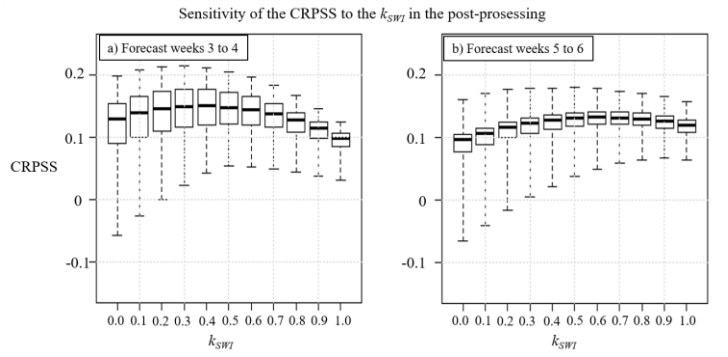


Figure 5. Sensitivity of the expected CRPSS of the ECMWF surface temperature reforecasts to the k_{SWI} ranging from 0.0 to 1.0 in forecast weeks 3–4 (a) and 5–6 (b). The black boxes show the lower and upper quartiles, and the whiskers illustrate the extremes of the November-February mean CRPSSs of all the grid points in Northern Europe.

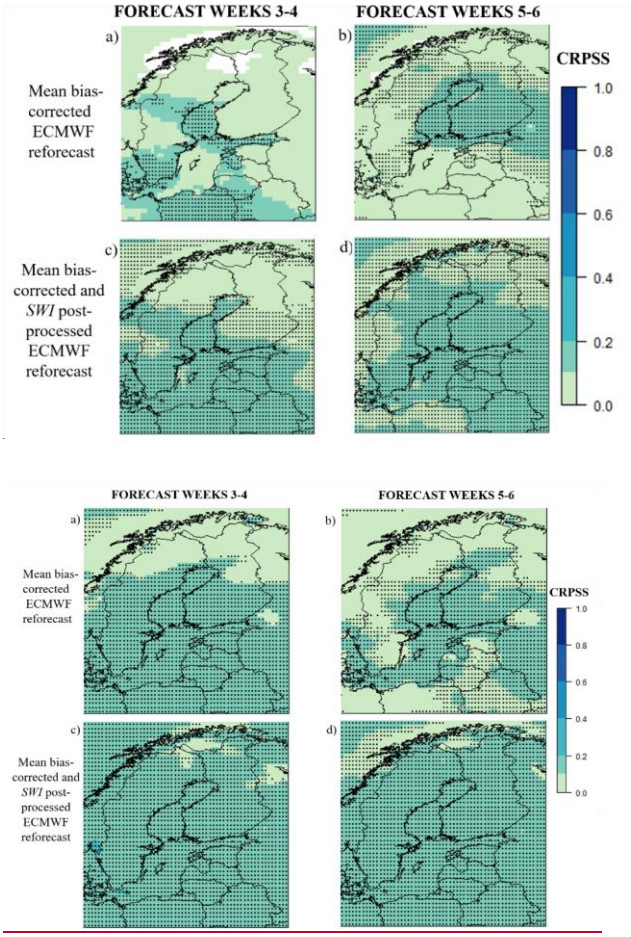


Figure 6. Expected CRPSS of forecast weeks 3–4 and 5–6 of the ECMWF's mean temperature reforecasts for November–February 1997–2016 after mean bias-correction (a-b) and after both mean bias-correction and the *SWI* based post-processing (c-d). ERA-Interim climatology of 1981–2010 was used as the reference. The dotted areas represent the 95% level of confidence that the CRPSS is above zero.