Stratospheric Modulation of Arctic Oscillation Extremes as Represented by Extended-Range Ensemble Forecasts

Jonas Spaeth and Thomas Birner

Meteorological Institute Munich, Ludwig-Maximilians-University, Munich, Germany

Correspondence: Jonas Spaeth (jonas.spaeth@physik.uni-muenchen.de)

Abstract. The Arctic Oscillation (AO) describes a seesaw pattern of variations in atmospheric mass over the polar cap. It is by now well established that the AO pattern is in part determined by the state of the stratosphere. In particular, sudden stratospheric warmings (SSWs) are known to nudge the tropospheric circulation toward a more negative phase of the AO, which is associated with a more equatorward shifted jet and enhanced likelihood for blocking and cold air outbreaks in mid-latitudes. SSWs are also thought to contribute to the occurrence of extreme AO events. However, statistically robust results about such extremes are difficult to obtain from observations or meteorological (re-)analyses due to the limited sample size of SSW events in the observational record (roughly 6 SSWs per decade). Here we exploit a large set of extended-range ensemble forecasts within the subseasonal-to-seasonal (S2S) framework to obtain an improved characterization of the modulation of AO extremes due to stratosphere-troposphere coupling. Specifically, we greatly boost the sample size of stratospheric events by using potential SSWs (p-SSWs), i.e., SSWs that are predicted to occur in individual forecast ensemble members regardless of whether they actually occurred in the real atmosphere. For example, for the ECMWF S2S ensemble this gives us a total of 6101 p-SSW events for the period 1997-2021.

A standard lag-composite analysis around these p-SSWs validates our approach, i.e., the associated composite evolution of stratosphere-troposphere coupling matches the known evolution based on reanalyses data around real SSW events. Our statistical analyses further reveal that following p-SSWs, relative to climatology: 1) persistently negative AO states (> 1 week duration) are 16% more likely, 2) the likelihood for extremely negative AO states ($< -3\sigma$) is enhanced by at least 35%, while that for extremely positive AO states ($> +3\sigma$) is reduced to almost zero, 3) ~~a p-SSW preceding an~~ approximately 50% of extremely negative AO ~~state within 4 weeks is causal for this AO extreme (in a statistical sense) up to a degree of 27%~~states that follow SSWs may be attributed to the SSW, whereas about one quarter of all extremely negative AO states during winter may be attributed to SSWs. A corresponding analysis relative to strong stratospheric vortex events reveals similar insights into the stratospheric modulation of positive AO extremes.

1 Introduction

Day-to-day variability of the northern extratropical hemispheric-scale circulation during winter is dominated by the so-called Northern Annular Mode (NAM, **?**). The surface manifestation of the NAM is often referred to as Arctic Oscillation (AO). This variability pattern primarily describes fluctuations of atmospheric mass over the polar cap with associated opposite fluctuations on its equatorward flank. In its positive phase the AO corresponds to ~~increased~~ decreased mass over the polar cap with associated strengthened pressure gradient across mid-latitudes that goes along with a stronger polar-front/eddy-driven jet that is shifted poleward and more zonally aligned. Likewise, in its negative phase the jet is weakened, shifted equatorward and often more meriodonally distorted.

Although a single index cannot represent the entire extratropical weather, it indicates tendencies towards certain weather patterns, which in turn can also have strong local effects. Especially AO values that deviate considerably from 0 (the climatological mean) are rare, by construction, and can often be associated with strong *local* weather extremes (**?**): For instance, the daily AO index was around -2.5 in winter 2009/10, which was accompanied by record cold snaps and snow fall over large parts of the United States, Europe and East Asia (**?**). In winter 2019/20, extreme storminess over Central Europe occurred during a highly positive AO phase with wind gusts of up to 177 km/h being recorded over Germany (**?**). Furthermore, **?** report increased likelihood of Siberian wildfires following positive AO periods in February and March.

The AO can also be influenced by "external" weather patterns and one prominent teleconnection exists between the AO and the stratospheric polar vortex. The latter describes a strong westerly wind band around 60°N extending over 10 hPa, which forms every year in winter (**?**). Numerous studies show that, on average, a very strong polar vortex (SPV) is associated with a strengthened circumpolar flow in the troposphere - as indicated by a positive AO index (e.g., **???**). The reverse is true for a weak polar vortex, with such events being a special case: The breaking of planetary waves in the stratosphere and the associated westward forcing can lead to a complete breakdown of the polar vortex. In these cases, the zonal mean zonal wind reverses and the climatologically dominant westerly winds are replaced by weak or moderate easterlies. During the vortex disruption, air masses converge in the center of the vortex and are forced to sink. The accompanying strong and rapid adiabatic heating is the reason that such extreme weak vortex events are called sudden stratospheric warmings (SSWs, **?**). SSWs are observed about 6 times per decade and are, as described previously, associated with a negative AO index on average. On synoptic scales, SSWs have also been tight to subsequently favored occurrence of certain weather regimes over the North Atlantic (**?**) and over North America (**?**).

Consistent with the local implications of a negative AO index, SSWs can for example lead to cold spells in Northern Europe and increased storminess over Southern Europe (**?**, and references herein). Whether it is generally valid that SSWs ~~,~~and also strong polar vortex events ~~,~~lead to a subsequently more likely occurrence of AO extremes (and associated local extremes) is difficult to analyze because the statistical links are weak in each case, i.e., not each SSW/SPV event is followed by an AO extreme. Therefore, a very large sample of SSW and SPV events are needed to quantify the subsequent risk increase of AO extremes. However, reanalyses data only cover about 40-70 years, depending on the data set, and thus about 30-40 SSWs - too few to robustly determine conditional probabilities (e.g., given a stratospheric extreme event, how likely is a following tropospheric extreme event).

In order to allow for analyses of larger event sample sizes, past studies have used, for example, idealized model simulations (e.g., **??**). Even though such models have proven to be useful to develop a qualitative and conceptual picture, they often show a weaker tropospheric response to stratospheric events compared to observational data (**?**). In this study, we aim to improve the characterization of coupled stratospheric and tropospheric circulation extremes using operational, state-of-the-art, extended-range forecasts. Relatively large ensembles, frequent model initializations and the generation of hindcasts ~~allows~~ allow us to analyze a large set of predicted SSWs and SPV events (p-SSWs/ p-SPVs, see discussion in section 2). Although the vast majority of these p-SSWs did not materialize in the real atmosphere we show that they nevertheless provide reliable statistical information about stratosphere-troposphere coupling. Our analyses implicitly assume that each ensemble member corresponds to a possible real-atmospheric evolution. The diagnosed p-SSWs include false alarm events (see, e.g., **?**), which we assume are based on the same underlying physics as those SSWs that occurred in the real atmosphere. Furthermore, the individual evolution (related to forecast score) is arguably not relevant for statistical characterizations of circulation anomalies.

The analysis is thus based on the assumption that the forecast models simulate the observed variability of the AO sufficiently well, including its modulation due to stratospheric variability. High-top models, in particular, show realistic stratosphere-troposphere coupling (**??**). However, due to the small sample size of observed events, it is generally difficult to conclude whether any discrepancies between model and observational data are due to model or sampling errors. For this study, we will show that the models agree with observations in established diagnostics that can be robustly derived from reanalyses, including, e.g., the frequency of SSWs, their seasonality and their average impact on the subsequent AO. Although our quantitative statistical analyses cannot be compared directly to observational data, they may be considered as best estimate given the currently available observational record and modeling capabilities.

We focus on following research questions:

1. By how much do stratospheric polar vortex extremes increase the probability of persistently positive or negative AO phases?

2. By how much do stratospheric polar vortex extremes increase the probability of subsequent AO extremes?

3. ~~How often can stratospheric polar vortex extremes be considered causal~~[1] ~~for subsequent AO~~ What fraction of AO extremes may be attributed to preceding stratospheric extremes?

The paper is organized as follows: Section 2 provides an overview of the extended-range forecasts used in this study. Section 3 defines stratospheric and tropospheric circulation extremes and presents basic event statistics. For SSWs, we validate in section 4 that the predicted events agree, in well-known diagnostics, with events that are identified in reanalysis data. This motivates section 5, where the probability of AO extremes following predicted SSWs is analyzed. Conversely, section 6 shows how often predicted AO extremes are preceded ~~(and caused)~~ by predicted SSWs and how many AO extremes may be attributed to preceding SSWs. Section 7 reveals in a similar fashion the statistical relation between predicted strong polar vortex events and predicted positive AO extremes, before the key findings are discussed and summarized in section 8.

2 Description of extended-range ensemble forecasts

The subseasonal to seasonal (S2S) prediction project (**?**) provides a collection of extended-range (up to 60 days lead time) ensemble forecasts from different weather services. Forecasts differ in terms of model specifications (e.g., spatial resolution,

[1] ~~in a statistical sense, see section 6 for a discussion~~

Table 1. Dataset specifications.

	S2S ECMWF	S2S UKMO	ERA5
Type	Forecast	Forecast	Reanalysis
Vertical Res.	L91	L85	L137
Time Range	d 0-46	d 0-60	1979-2021
Realtime	51 member, 2 inits / week	4 member, daily inits	-
Hindcast	11 member, 2 inits / week, past 20y	7 member, 4 inits / month, ~~1993-2016~~ 1993-2015	-
# Realtime Ens. Used	114	396	-
# Hindcast Ens. Used	2280	1173	-
# Individual Model Runs	30894	9795	-

parameterizations, maximum lead time). All forecast systems create hindcasts in addition to the realtime forecasts in order to calibrate the forecasts and to allow the construction of the model's climatology. For our application, the most relevant demand is an accurate representation of the stratosphere and in particular of stratosphere-troposphere coupling. Furthermore, a forecast model with a large number of hindcasts is beneficial, because it allows for more robust analyses by including multiple past years. Lastly, a large maximum lead time is needed as we want to identify extreme events in the forecasts and are then also interested in the time periods before and after the event.

We choose to use ECMWF and UKMO forecasts for this study, as these models best meet the above listed requirements. Importantly, both models have been demonstrated in previous studies to have a realistic representation of stratosphere-troposphere coupling (**??**).

For the decision on which initialization dates to use for the analyses, a trade-off has to be made between having as large a sample as possible and the fact that the forecast models are updated about every 1-3 years. Since late 2016, the ECMWF model (CY43R1) has been running at a higher horizontal resolution. Therefore, to avoid mixing forecasts before and after 2016, forecasts from winter 2017/18 up to and including 2020/21 are analyzed. Note that a minor model version change occurred in 2019, where initial conditions for the hindcasts are then obtained from ERA5 instead of ERA-Interim. However, we do not expect this to be a major limitation for our analyses, as we are mostly interested in the overall statistical behavior of stratosphere-troposphere coupling, as opposed to single forecast performance.

We focus on Northern winter dynamics by analyzing forecasts initialized between mid-November (11/16) and end of February (02/22). For the four winter seasons, the ECMWF model thus features 114 real-time ensemble forecasts of 51 members each and 2 280 ensemble hindcasts of 11 members each. This results in a total of 30 894 individual model runs, all of which we refer to as "forecasts" for simplicity. For consistency, UKMO forecasts are used from the same initialization period, leading to 9 795 forecasts available for this model. A summary of the key specifications of the forecasts is given in Table 1, along with details of the ERA5 data (**?**) used.

3 Event statistics of stratospheric and tropospheric circulation extremes

3.1 Data sets and overall methodology

Each of the forecasts from the total set of 30 894 ECMWF forecasts provides a 47-day time series of the evolution of the atmosphere (UKMO: 61 days). In this study, we define specific events and then scan each forecast for the occurrence of such an event. If there are multiple events of one event type within one forecast, only the first event is used. Note that, by definition, all identified events are predicted events, but each may or may not actually occur in the real atmosphere. To highlight this aspect, and to avoid confusion with actual real-atmospheric events, ~~we will refer to~~ the events identified in the forecasts may be denoted with a "p"-prefix~~in the following~~, where "p" stands for "predicted" (alternatively, it could be thought of as "potential" for some aspects). In this study, all event composites and computed probabilities refer to predicted events.

For both datasets, ECMWF and UKMO, all individual forecast runs are treated equally and independently. This assumption is violated especially for forecasts belonging to the same ensemble. In fact, at initialization time these forecasts agree *entirely* except for ensemble perturbations. The individual members diverge from each other only with increasing lead time, when the predictability of the atmospheric flow gradually decreases. For this reason, we analyze only those events that occur at or after a forecast lead time of 10 days. It is assumed that initial condition-memory has sufficiently reduced by this point so that no two individual forecasts fully match, and the same is thus true for the evolution of the identified events. This ensures a degree of statistical independence. The use of hindcasts further guarantees sampling of different boundary conditions, such as due to the El-Niño-Southern-Oscillation, the Maddan-Julian-Osciallation or sea ice variations.

Furthermore, it is ensured that for each identified event both negative and positive lags can be considered. Due to the finite maximum lead time of each forecast, this demand is generally limited. For a predicted event that occurs early in the forecast (but after 10 days at the earliest), only a short period before the event can be examined, and the reverse is true for an event that occurs shortly before the end of the forecast. Therefore, to ensure a minimum common lag time that can be analyzed, events are additionally required to occur no later than 10 days before the end of the forecast. Consequently, events are allowed to occur between day 10 and 36 for ECMWF forecasts and between day 10 and 50 for UKMO forecasts. Thus, for all events, the lag period ±10 days can be examined, but with increasingly longer positive and negative lag times, fewer and fewer events contribute to the composite.

Extreme events are defined that refer to exceptional anomalies in the stratospheric and tropospheric circulation, respectively. As a measure of the strength of the stratospheric polar vortex we use the zonally averaged zonal wind at 10hPa at 60°N, hereafter referred to as u60.

3.2 Predicted SSWs

We define Sudden Stratospheric Warmings (p-SSWs), as days when u60 transitions from positive to negative, i.e., the polar vortex breaks down. As explained above, we do not include p-SSWs predicted within the first 10 days after forecast initialization. However, for p-SSWs, u60 is required to be solely positive within these first 10 days to ensure an intact westerly polar

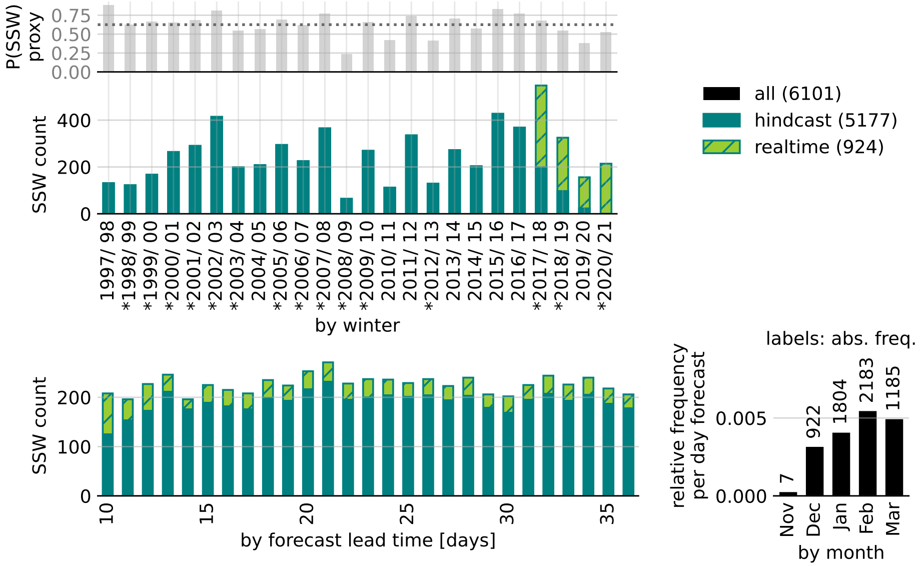

Figure 1. Distribution of analyzed p-SSWs in ~~the~~ ECMWF forecasts~~grouped by winter~~. Absolute event counts (center left) and seasonal probability proxy (top left), grouped by winter. Asterisks denote years with real atmosphere SSWs (**?**). Grouped by forecast lead time (bottom left) and by month (bottom right).

vortex at the start of the forecast. Following this event definition, we identify 6 101 p-SSWs in the ECMWF and 2 716 p-SSWs in the UKMO model.

~~As a sensitivity test we also applied our analyses to~~

Moreover, the analyses were repeated with a modified event definition, which we call *dynamical SSWs*, in order to investigate potential sensitivities. Dynamical SSWs were defined as a subset of ~~predicted *dynamical* SSWs(defined by a SSW where, in addition~~SSWs, where in addition to the sign change, u60 ~~drops~~ is required to drop at least 20 ms^{-1} averaged over -5 to $+5$ days lag relative to the SSW central date~~). The results agree with some surface signatures being enhanced, but as differences to p-SSWs are only small, we here restrict the analyses to p-SSWs. Furthermore, previous literature has suggested that polar vortex displacement events could be predictable at slightly longer lead times than displacement events. However, due to small sample sizes, the differences are generally not statistical significance and we therefore do~~. Thereby, this event definition forms the intersection between SSWs (following **?**) and sudden stratospheric deceleration events (following **?**, ensuring a rapid deceleration around th . Our results reveal only modest quantitative differences between SSWs and dynamical SSWs and we therefore focus on SSWs only, to allow better comparison with other studies.

~~—~~In Figure 1 we provide an overview about the distribution of ECMWF p-SSWs as a function of the year, forecast lead time ~~and~~ calendar month (see Fig. S1 for a corresponding analysis of UKMO forecasts). p-SSWs are found for all winter seasons considered~~(top left panel).~~

. Absolute numbers are presented to show which winter seasons contribute how many events to the analysis. Due to the realtime-hindcast-setup, the number of underlying forecasts varies across winter seasons. Therefore, we additionally provide a proxy for the SSW probability per winter season to illustrate inter-annual variability (see appendix B for details).

The largest number of events is identified in the winter season 2017/18, which ~~has~~ includes also the most forecasts (realtime 2017/18 plus hindcasts related to initializations from 2018/19 to 2020/21). Different factors lead to a highly varying number of events between the different years. These include internal dynamic variability, a slightly varying number of underlying forecasts, due to the realtime/ hindcast prediction setup, and the varying number of events per winter due to the evolution of the polar vortex of the real atmosphere in the respective winter, which determines the initial conditions of the forecasts.

A forecast that is initialized with a strong polar vortex tends to maintain a strong polar vortex and produces fewer SSWs compared to a forecast with an initially weak polar vortex. Moreover, forecasts that do not start with ten consecutive days of positive u60 are discarded by default. Thus, if the polar vortex in the real atmosphere is already easterly at the initialization time or is predicted to become easterly within the quasi-deterministic forecast range of ten days, such forecasts will not contribute any events to the analysis. This can be illustrated by the example of the ~~2009'th~~ 2009th SSW (24 January 2009, see **?**). The event had low predictability at lead times longer than 8 days (**?**). ~~Moreover, the eventwas generally very rare due to the polar vortex being exceptionally strong~~ Before the event, between end of December 2008 and mid-January ~~2009.~~ 2009, the polar vortex was exceptionally strong, leading to an only marginal SSW probability in the forecasts and suggesting that the event itself was unlikely given the prevailing dynamics. As a result, 2008/09 shows the lowest number of SSWs: In the first winter half up to initialization dates around mid-January, hardly any events were predicted due to the relatively strong polar vortex. Later, forecasts predicting the real atmosphere SSW only did so at less than +10 day lead time, such that those events are discarded. Later initializations up to mid-February are excluded, because these do not predict persistent positive u60 within the first 10 days lead time, due to the preceding SSW. As a result, winter season 2008/ 09 contributes only 64 (UKMO: 22) p-SSWs to the analysis, and at 23% (UKMO: 41%), the approximated SSW probability is the smallest in the period considered.

Based on the average number of 226 events per day lead time in the ECMWF model (cf. bottom left panel in Fig. 1), we estimate the ~~corresponding probability of at least one SSW per winter season (≈ 135 days from~~ probability of a SSW between mid-November ~~to~~ and end of March~~):~~

$$P(\text{SSW}) = 1 - \underbrace{\underbrace{\left(1 - \frac{226}{30894}\right)}_{\text{no SSW}}^{135}}_{\text{for 135 consecutive days}} = 63.3\%/\text{ winter}$$

, which yields 63% (see appendix B for details). This is consistent with the number of observed SSWs in reanalyses that is roughly 6 per decade (**?**).

While the rate of events per forecast day fluctuates only weakly in the ECMWF model, it moderately increases with lead time in the UKMO model (Fig. S1, bottom left panel). One might expect this to be due to the longer maximum lead time of the UKMO model (+60 days) compared to the ECMWF model (+46 days), which may allow more final-warming-like events.

However, we find that the trend is still apparent when all forecasts initialized in February are excluded from the analyses (not shown).

Consistent with reanalyses (e.g, **?**) and across both, the ECMWF and the UKMO model, the p-SSW frequency shows a maximum in February (bottom right panel in Fig. 1). However, **?** find leadtime-dependent inconsistencies in the seasonal distribution of SSW probability compared to the observational record.

3.3 Predicted strong vortex events

Past literature has identified stratosphere-troposphere coupling not only following SSWs, but also following strong polar vortex events (SPVs, e.g., **?**). However, the definition of a single event in these cases is somewhat more ambiguous, as there is no dynamically motivated threshold for u60, compared to 0 ms^{-1} for SSWs. In addition, the dynamical changes in cases of a strong polar vortex are generally less abrupt, making it harder to pin down one particular central event day. For these reasons, we focus mainly on SSWs in this paper, however, we also provide a summary of the key results for SPV analyses in section 7. In these analyses, p-SPVs are defined as the first day on which u60 exceeds a threshold that, based on percentiles, represents the "opposite" of the SSW threshold of 0 ms^{-1}. Depending on the model's climatology, this threshold describes approximately the 91st percentile of the u60 distribution and is around 47 ms^{-1}.

3.4 Predicted ~~NAM1000~~ AO events

In the troposphere, ~~extreme events are defined~~ we define extreme events based on the Arctic Oscillation Index (short: AO; equivalent to the Northern Annular Mode Index at 1000 hPa~~(~~, short: NAM1000~~, or AO[1]). This~~). The index is calculated by first area-weighting the geopotential field between 65 and 90°N by the cosine of latitude and then averaging over the entire polar cap. The ~~NAM~~ AO index then is the negative standardized anomaly of the obtained quantity. For technical details about the deseasonalization via the hindcasts, the reader is referred to appendix A. The positive phase of the ~~NAM1000~~ AO describes a negative geopotential anomaly over the polar cap and a thereby induced enhanced circumpolar westerly circulation. Conversely, a negative ~~NAM~~ AO reflects a weaker westerly circulation, which is typically associated with a southward shift of the jet that is also zonally more distorted.

We define tropospheric extreme events as the first day when the ~~NAM1000 falls below -3 (p-NAM1000$^-$ extreme~~AO falls below a certain negative threshold (e.g., AO^{-3} corresponds to $AO < -3$) or exceeds ~~$+3$ (p-NAM1000$^+$ extreme~~a certain positive threshold (e.g., AO^{+3} corresponds to $AO > +3$). After testing different thresholds, we opt for ~~a threshold of~~ thresholds of up to 3 standard deviations ~~because it~~ which represents a tradeoff between severity of event and sufficiently large resulting sample sizes.

3.5 Conditional probabilities of polar vortex and AO extremes

[1] ~~We will use the notation "(p-)NAM1000" where we explain and refer to technical details. Due to better readability and more widespread usage in other literature, we use the term "AO" where we make generalized statements and in the conclusions. However, we note that both terms are interchangeable.~~

In this study, conditional probabilities are computed to quantify the modulated likelihood of AO extremes under the presence or absence of preceding stratospheric extremes. For example, we expect the probability of at least one AO^- extreme during a given time period to be higher if that time period follows a SSW compared to the case that it does not follow a SSW. This is somewhat akin to the situation in climate attribution science, where one aims to quantify the increased risk of an extreme event due to anthropogenic climate change (e.g., ?), or to the situation in epidemiology, where one aims to quantify the increased risk of contracting a disease given an exposure to a particular factor (e.g., smoking in the case of lung cancer; ?). In such situations one may quantify the additional risk due to the exposure based on the so-called relative risk increase (RRI):

$$\mathrm{RRI} = \frac{\text{risk among the exposed}}{\text{risk among the unexposed}} - 1$$

In climate attribution science "exposure" may be thought of as "under the influence of anthropogenic climate change", whereas lack of exposure (the condition in the denominator) may be thought of as "without the influence of climate change" (e.g., based on pre-industrial control climate). In our case of stratosphere-troposphere coupling exposure may be thought of as "given that a stratospheric extreme occurred". However, lack of exposure has to be evaluated with care. For example, assume that a given day t_0 fulfills the condition of "no stratospheric extreme" and an AO extreme occurs within a given period following t_0. This AO extreme cannot necessarily be considered "unexposed" as a stratospheric extreme may have occurred between t_0 and the date of the AO extreme. For our analyses that evaluate the increased probability of an AO extreme following a stratospheric extreme event we therefore adopt a modified version of RRI, where we replace the denominator with the risk of AO extreme occurrence for the population (i.e., including both exposed and unexposed). To avoid confusion we will refer to this modified RRI simply as "relative probability increase" (see section 5).

One way to circumvent the above discussed issue of conditioning onto "unexposed" is to swap the conditioning. That is, we may condition onto the occurrence of an AO extreme and evaluate the probability that a given preceding time period showed at least one day with stratospheric extreme occurrence – in this case the AO extreme is considered to be "exposed". Likewise, if the preceding time period shows no occurrence of stratospheric extreme, the AO extreme is considered to be "unexposed". Using Bayes theorem this allows us to estimate the fraction of attributable risk (FAR) of AO extremes due to a preceding stratospheric extreme. We will distinguish FAR among the exposed and among the population (see section 6).

Detailed mathematical definitions of relative probability increase, attributable risk among the exposed and among the population will be introduced in the respective sections. Nevertheless, we here provide an overview table about the event definitions that will be used (Tab. 2).

4 Evaluation of stratosphere-troposphere coupling based on predicted SSWs

To provide a baseline for our more detailed statistical analyses in the following sections, we first evaluate the general behavior of stratosphere-troposphere coupling based on p-SSW events in the S2S data. To do so we focus on the lag-composite evolution of the ~~NAM1000~~ AO index relative to p-SSWs compared to real-atmospheric SSWs from ERA5. In addition, we show the NAM

Table 2. Definitions for (conditional) predicted SSW and AO events. Subscript wt is short for "within time t". AO events can be negative (AO^-) or positive (AO^+) and may refer to a prescribed threshold, i.e., AO^{-3}_{wt} or AO^{+3}_{wt} correspond to "at least one day below -3 or above $+3$ within time t".

Event	**Description**
AO	probability that any day shows an AO extreme
AO_{wt}	probability that any period of time t shows at least one AO extreme
$AO_{wt} \mid SSW$	given a SSW, probability of at least one AO extreme within subsequent time t
SSW_{wt}	probability that any period of time t shows at least one SSW event
$\neg SSW_{wt}$	probability that any period of time t shows no SSW event
$SSW_{wt} \mid AO$	given an AO extreme, probability of at least one day with u60 <0 within preceding period of time t
$\neg SSW_{wt} \mid AO$	given an AO extreme, probability of no day with u60 <0 within preceding period of time t
$AO \mid SSW_{wt}$	given a preceding period of time t where at least one day with u60<0, probability of AO extreme on day afterwards
$AO \mid \neg SSW_{wt}$	given a preceding period of time t where no day with u60<0, probability of AO extreme on day afterwards

index at 200hPa (short: NAM200) because the lower stratosphere has been found to play an important role in stratosphere-troposphere coupling (e.g, **??**).

Figure 2 shows the evolution of u60 (top), NAM200 (center) and ~~NAM1000~~ AO (bottom) during SSWs, averaged over all events, separately for ECMWF and UKMO. In addition to the composite mean, the 33rd to 66th percentiles across all ECMWF events on the respective lag day are shown. By construction, 100% of all events (ECMWF: 6 101, UKMO: 2 716) contribute to lag days ±10. For larger positive or negative lags, some forecasts have reached their maximum forecast lead time or have not yet been initialized. Therefore, the number of events drops off, which makes the statistics less robust: For the ECMWF model, the number of contributing events falls below 20% for lags smaller than -31 and larger than +31 days (UKMO: smaller than -44 and larger than +39 days).

By construction, u60 transitions from westerly to easterly at lag 0. Anomalies of u60 are slightly positive ahead of -14 days lag, which we interpret as an indication for vortex preconditioning (**???**). The anomalies become negative within the second week prior to the event central date. The largest average negative anomalies occur only few days after the event central day (lag +2 days: -6ms^{-1}). Afterwards, the vortex reestablishes and the average anomalies reach zero again after approximately 35 days. Consistent with, e.g., **?**, both NAM200 and ~~NAM1000~~ AO are negative following the event. The shift of the NAM200 happens earlier (at lag day -11) and the timing aligns well with the weakening of the polar vortex at 10hPa. The NAM200 anomaly is also more pronounced (≈ -0.5) compared to the ~~NAM1000~~ AO (≈ -0.3). Interestingly, the ~~NAM1000~~ AO distribution is slightly shifted toward positive values in the week prior to the central date, which is also robust for other diagnostics like the 10th, 30th, 70th and 90th percentiles (not shown). At long positive lag times, the NAM indices at 200 and 1000hPa are still negative (ECMWF: lag +36 days, UKMO: lag +51 days), but the trend goes to weaker ~~positive~~ negative values again.

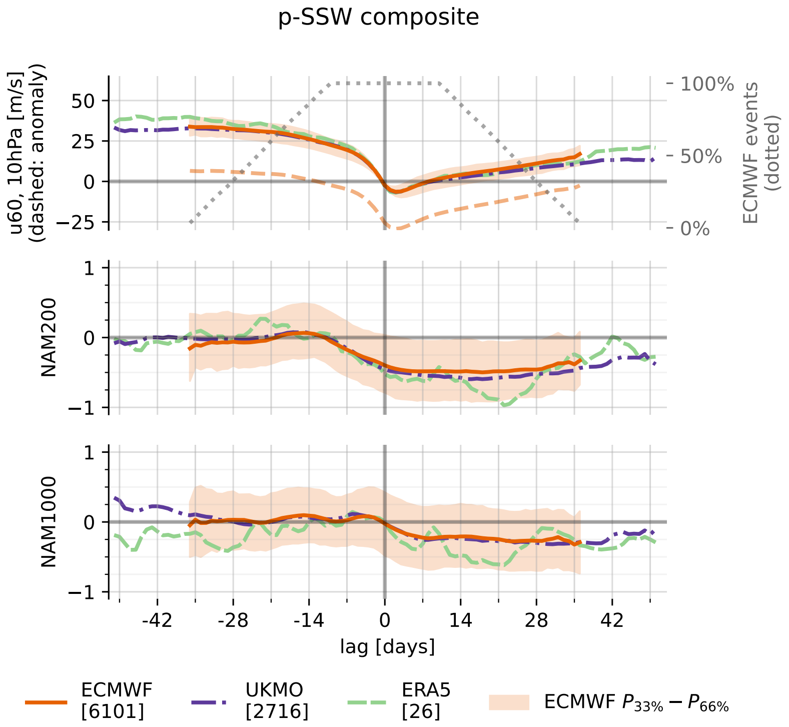

Figure 2. Lagged composite evolution of u60 (top panel), NAM200 (middle panel) and NAM1000 (=AO, bottom panel) relative to p-SSWs (ECMWF, UKMO) and SSWs (ERA5). It is presented the mean across all ECMWF events (orange, solid), the 33rd to 66th percentiles across all ECMWF events (orange, shaded), the mean of all UKMO events (purple, dash-dotted) and the mean across all ERA5 events (green, dashed). In the top panel further denoted are the average u60 anomalies (orange, dashed) and the relative number of contributing events to the composite in the ECMWF model (gray, dotted). Square brackets denote the total number of events, for each dataset.

Overall, the results are in agreement with ERA5 and previous literature and especially the evolution of u60 is remarkably similar. The negative NAM response at 200hPa and 1000hPa seems to be slightly stronger in the reanalysis, however, it is also noisier due to the smaller sample size.

5 Predicted AO extremes following predicted SSWs

In the following, we will exploit the larger available sample size of p-SSW events to diagnose and quantify whether the shift of the average AO index towards negative values is caused by 1) more persistent negative AO phases and/or 2) an increased probability ~~for~~ of AO^{-} extremes.

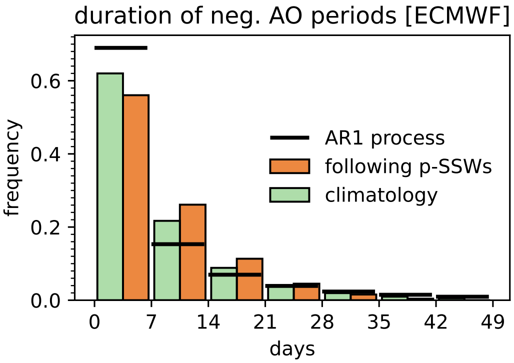

Figure 3. Histogram of the duration of negative ~~NAM1000~~ AO periods, quantified by the number of consecutive days of ~~NAM1000~~ AO < 0 and binned by 7-day chunks. Periods following ECMWF p-SSWs (orange bars, right half) are compared to the ECMWF model's climatology (green bars, left half) and a random first order auto-regressive model of the same 1-day-lag-autocorrelation as the ~~NAM1000~~ AO in ERA5 (black, horizontal lines).

5.1 Persistence of negative AO phases

Figure 3 presents a histogram of the duration of predicted negative ~~NAM1000~~ AO phases in the ECMWF ~~S2S~~ model, binned into 7 day chunks. The duration is defined as the number of consecutive days with negative ~~NAM1000~~AO. The climatology serves as a reference including all 30894 ECMWF ~~forecast~~ forecasts used for this study. With approximately 62%, most phases of negative ~~NAM1000~~ AO are shorter than 8 days. As another reference, a first order autoregressive model was set up with zero mean and standard deviation of 1, which may serve as a baseline. Its 1-day-autocorrelation is chosen to match the ERA5 ~~NAM1000~~ AO timeseries and for robustness, it is estimated by averaging the lag-1-autocorrelation and the square-root of the lag-2-autocorrelation, yielding 0.91. ~~It turns out that the NAM1000~~ The respective AO climatologies in ECWMF (S2S) and ERA5 agree very well (not shown). However, the AO climatology shows short negative phases (≤ 7 days) less often and long negatives phases (≥ 8 days) more often compared to the AR1 process, indicating an AR1 process cannot reproduce the observed AO variability.

~~This is an indication for the NAM1000 index having a slightly longer decorrelation timescale in the S2S model compared to ERA5; which apparently also overwhelms the effect of negative NAM1000 periods being cut off by the end of the forecast which introduces a bias towards shorter negative periods.~~

In addition, the diagnostic is presented for periods following p-SSWs. Here, the ~~NAM1000~~ AO index is analyzed between lag day +1 relative to the event date and the maximum available lag time, which ranges between +10 and +36 days, depending the forecast lead time when the event happens. Similar to the reference climatology, this diagnostic also underestimates the occurrence of long negative ~~NAM1000~~ AO periods as the forecasts have finite maximum lead time. Nevertheless, periods following SSWs show a reduced frequency of shorter and an increased frequency of longer negative ~~NAM1000~~ AO periods, compared to the climatology (and thus also to the AR1 process): For instance, 38% of negative ~~NAM1000~~ AO periods are

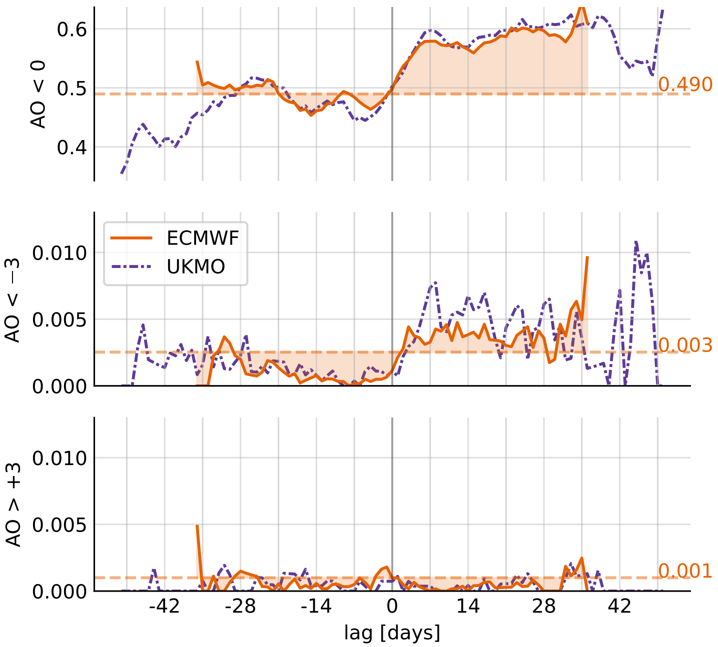

Figure 4. Daily ~~probability~~ probabilities of ~~p-NAM1000~~ AO<0 (top panel), ~~p-NAM1000~~ AO< -3 (middle panel) and ~~p-NAM1000~~ AO$>+3$ (bottom panel) relative to p-SSWs, quantified by the fraction of events fulfilling the respective condition~~. The statistics are presented~~, separately for ECMWF ~~events~~ (orange, solid) and UKMO ~~events~~ (purple, dash-dotted)~~and~~. Day 0 corresponds to the p-SSW central date. In addition, probabilities are compared to the corresponding daily ECMWF climatology (dashed horizontal lines). ~~For each lag day, the number of events fulfilling the respective condition has been normalized by the total number of events contributing to the composite on that lag day.~~

longer than 7 days in the climatology, whereas this probability rises to 44% following p-SSWs, which corresponds to a relative increase of 16%~~(UKMO: also 16% ,~~.

Sampling uncertainties turn out to be negligible within 95% confidence intervals. A similar analysis based on UKMO data shows very good quantitative agreement (not shown), which further confirms the robustness of the results.

5.2 Modulated probability of AO extremes

~~We now focus on p-NAM1000 extreme events and analyze to what extent p-SSWs contribute to an increased probability for such events. As the NAM1000 distribution shifts at positive lag times~~ It is known that SSWs shift the subsequent AO distribution (see Fig. 2)~~, also the daily statistical probability of extreme NAM1000 values changes. This effect is quantified in~~. This also implies an increased daily probability of negative and a reduced probability of positive AO extremes compared to their respective climatological probabilities. Fig. 4 ~~. First, based on all available forecasts, the climatological likelihood is computed for~~ shows the probabilities of negative (<0), extremely negative (<-3) and extremely positive ($>+3$) ~~p-NAM1000 events on any random day ,~~ AO values on a particular lag day t relative to the SSW central date. Mathematically, these probabilities can be written as $P(AO\,|\,SSW)$. Per construction, lag day 0 describes the SSW central day. At each lag day, the probabilities are computed by normalizing the ~~days~~ number of events fulfilling the respective condition with the total number of available ~~forecast days. The resulting probability baseline for the ECMWF forecasts is 49% for negative, 0.3% for extremely negative and 0.1% for extremely positive events, where the asymmetry is due to the negative skewness of the NAM1000 distribution (0.13~~events at the respective lag day (which decreases for large positive and negative lags).

In addition, the overall daily probabilities of AO<0, AO<-3 and AO$>+3$ are presented, providing climatological baselines $P(AO)$, which are independent of lag time. In any forecast, AO events occur at each day with probabilities of about 49.0% for AO <0, about 0.3% for $AO<-3$ and about 0.1% for $AO>+3$. Asymmetry between positive and negative values arises from the AO distribution that is not perfectly Gaussian (skewness: -0.13).

The fraction of events in the p-SSW composite that have negative ~~NAM1000~~ AO values fluctuates around ~~50%~~ $P(AO^{-0}\,|\,SSW)=50\%$ at negative lags with only small deviations from the climatology. Within the first week following the event, this fraction increases and appears to saturate around 60%. Consequently, in the period following a p-SSW, a negative ~~NAM1000~~ AO is, at each day, approximately 50% more likely compared to a positive ~~NAM1000~~ AO (60% vs. 40%). The results are consistent between ECMWF and UKMO during the ±4 week period where the composites for both models consist of more than 30% of all events.

Extremely negative ~~NAM1000~~ AO values in the dataset appear with a climatological probability that is similar to what would be expected for a (one-sided) 3-sigma-event of a standard normal distribution (0.27%). At negative lags, they occur overall less frequent compared to climatology. In contrast, around lag 0, the probability increases and persists at ~~$\approx 0.40\%$~~ $P(AO^{-3}\,|\,SSW)\approx 0.40\%$ for more than four weeks. The increase appears to be larger in the UKMO model, however due to fewer events the diagnostic is also noisier. The fraction of events with extremely positive ~~NAM1000~~ AO values is smaller compared to climatology throughout the entire lag period. This is largely consistent between the models from ECMWF and UKMO. ERA5 (not shown) overall reveals higher probabilities of negative AO values following SSWs, $P(AO^{-0}\,|\,SSW)$. However, large uncertainties (95%-CI $\approx[45\%;85\%]$) in ERA5 make it difficult to distinguish whether observed differences arise from sampling errors in the reanalysis or from imperfect models. The ERA5 baseline probabilities of AO extremes are modestly lower compared to the S2S models[1] and not a single AO$^{\pm3}$ extreme event occurred within a four week period following a real atmosphere SSW, resulting in $P^{ERA5}(AO^{\pm3}\,|\,SSW)=0$, likely due to a very limited sample size.

An altered probability of extreme ~~NAM1000~~ AO events may be of higher socio-economic relevance than a small shift of the mean. However, the absolute daily probabilities of extremely negative ~~NAM1000~~ AO events are still small even though the relative increase ~~due to~~ given the p-SSWs is indeed considerable. In practice, the relevant question might not be how much the probability increases on only one specific day following a p-SSW. It may be more relevant to quantify the increased risk for an extreme ~~NAM1000~~ AO within a given time period ~~that is due to~~ following a p-SSW.

Figure 5 therefore shows the probability ~~P~~ of at least one ~~p-NAM1000^{-}~~ AO^{-3} extreme between day 1 and day t as a function of t. We compare the period following p-SSWs~~with the climatology of the UKMO and the ECMWF model~~, $P(AO^{-3}_{wt}\,|\,SSW)$

[1] Note that we have standardized the AO in ERA5 such that the inter-annual standard deviation is 1, similar to the deseasonalization that is applied to the S2S forecasts.

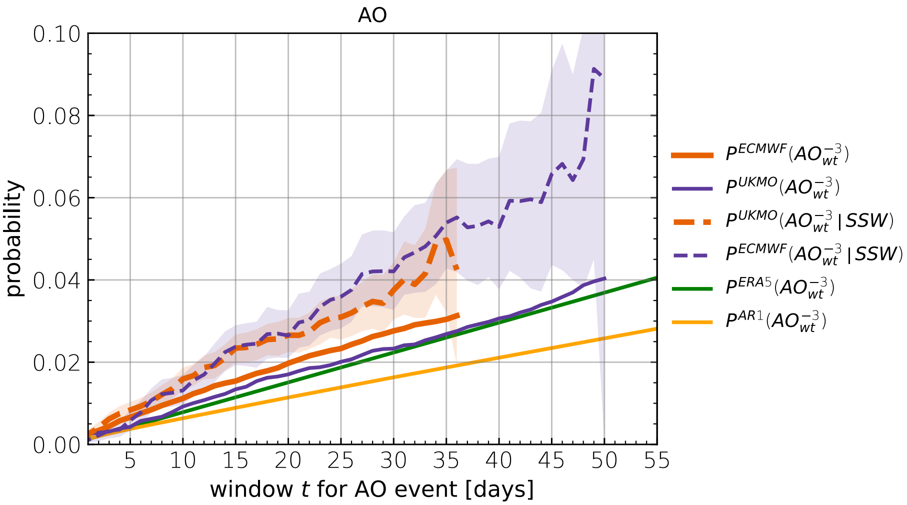

Figure 5. ~~Probability~~ Probabilities of at least one ~~day of p-NAM1000 < -3~~ AO^{-3} event within ~~days 1 to~~ a window of time t ~~, where periods~~ following p-SSWs (dashed, mean incl. 95%-confidence-interval) are compared to ~~the model's~~ climatology (solid~~+ shading: mean incl. 95%-confidence-interval~~), separately for ~~the~~ ECMWF (orange) and ~~the~~ UKMO (purple)~~model~~. In addition, the ~~climatology of~~ climatologies for ERA5 (~~solid,~~ green) and a random first-order auto-regressive model of the same 1-day-autocorrelation (yellow) are presented.

to the respective model climatologies, $P(AO^{-3})$, the ERA5 climatology and an AR1 process of the same autocorrelation as the ~~NAM1000~~ AO index in ERA5. ~~For the ECMWF and the~~ Confidence intervals were obtained for $P(AO_{wt}^{-3}\,|\,SSW)$ by bootstrap sampling all SSW events. For ECMWF and UKMO climatology, ~~the probability was sampled for all the forecasts~~ probabilities were sampled from lead time +10 days[2] to lead time +10+t days ~~. For the sampling, bootstrapping was applied where n random forecasts were picked and analyzed for p-NAM1000^{-} events, providing also an estimate for the 95% confidence interval (with n being the number of forecasts in the p-SSW composite). For~~ within all forecasts. Similarly, baseline probabilities of ERA5 and the AR1-process ~~, the probability is sampled~~ are obtained by sampling from all days t_0 of the time series to day $t_0 + t$, respectively.

~~The probability for ECMWF and UKMO p-SSWs is computed between lag day 1 of the respective composite until lag day~~

Clearly, all probabilities increase with t~~. This is realized by computing the counter-event, i.e.: $1 -$ "no p-SSW between day 1 and day t"[3]. Naturally for all datasets, as t increases, also the probability P increases, ~~, as the time window for finding at least one ~~NAM1000^{-}~~ AO^{-3} extreme gets wider.

However, with increasing t, also fewer events contribute to the composite due to the finite forecast lead time, leading to larger sampling errors. The results show that p-SSWs are consistently leading to an increased ~~integrated risk of extremely~~

[2] ~~day 10 is the first where~~ as we also start to search for p-SSWs at lead time day 10, however, this choice is arbitrary and the resulting climatology is not very sensitive to this choice

[3] ~~Special care must be taken when normalizing the events matching the condition by the total amount of events where the latter 1) must be subtracted by the number of events that have already matched the condition at earlier t's and 2) generally decreases for longer lag times.~~

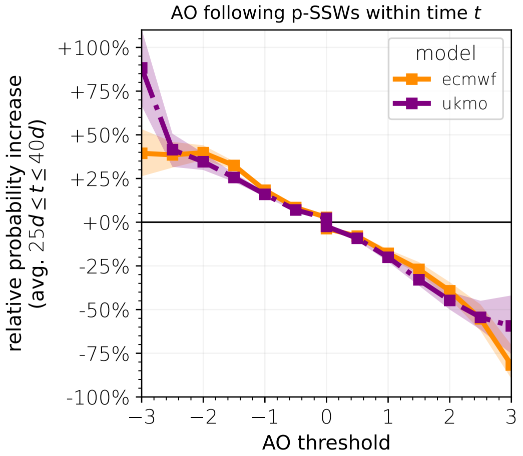

Figure 6. ~~Estimated probability~~ Probability increase ~~of~~ for at least one negative ~~p-NAM1000 extremes~~ (positive) p-AO extreme below (above) threshold following p-SSWs within a certain period t, relative to ~~the model's~~ climatology, averaged over 25d$\leq t \leq$ 40d, separately for ~~the~~ ECMWF (orange, solid) and ~~the~~ UKMO (purple, dash-dotted)~~model~~. ~~The estimate is obtained by dividing the curve for p-SSWs in Fig. 5 through the respective model's climatology and averaging over t. In Fig. 5, this is only shown for p-NAM1000 < -3; here, we present the diagnostic as a function of this p-NAM1000 threshold.~~

~~negative NAM1000~~ time-integrated risk of AO^{-3} events. For example, the probability in the ECMWF forecasts of at least one ~~NAM1000$^-$ extreme within 28~~ AO extreme within 30 days following the event is ~~3.4~~3.8%, compared to ~~2.6~~2.9% for its climatology. Overall, p-SSWs seem to affect the probability more in the UKMO model, as the probability following p-SSWs is higher and the climatological baseline is also lower compared to the ECMWF model. The baseline in ERA5 is slightly lower than in the ECMWF model, but agrees well with the UKMO climatology. All probabilities range considerably higher than the probability of a one-sided 3-sigma event for the AR1-process and as before, this is a result of the negative skewness of the ~~NAM1000~~ AO distribution.

Generally, all probabilities appear approximately linear in t, but it should be noted that the linear regime only holds for small enough t ~~,~~ as the probability will approach 1 and saturate in the limit of very large t. Furthermore, it is expected that for much larger t (which cannot be ~~displayed~~ evaluated here, due to the finite maximum forecast lead time), the effect of a p-SSW increasing the subsequent extreme ~~NAM1000~~AO$^-$ probability diminishes and the climatology will approach the one for p-SSWs.

~~It has been tested and verified that the results do not change significantly when forecasts containing p-SSWs are left out for the computation of the climatology, suggesting that p-SSWs do increase but not dominate the total number of p-NAM1000$^-$ extreme events.~~

~~We have shown that the time-integrated probability for~~ Based on the presented probabilities, the probability increase of at least one ~~day with NAM1000 < -3 is increased by a SSW. When displayed as~~ AO event within time t following SSWs can be determined *relative* to the climatological baseline:

$$\text{relative probability increase} = \frac{P(AO_{wt} \mid SSW)}{P(AO_{wt})} - 1$$

~~A~~ relative probability larger than 0 corresponds to an increase of AO probability following SSWs, while negative values ~~de~~scribe a probability decrease. The ratio is a function of the ~~period that is used to search for such p-NAM1000 events, the probability increase *relative to the climatological baseline* is roughly constant (e. g., 2% versus 1.5% $\hat{=} \times 1.33$ after 14 days, 3.4% versus 2.5% $\hat{=} \times 1.36$ after 28 days)~~length of the time window t (see supplement Fig. S2) and is assumed to approach 1 in ~~the~~ limit of large t, as the SSW influence becomes negligible. However, for medium time windows t that correspond to a typical ~~tim~~escale of stratosphere-troposphere coupling, the relative probability shows a wide plateau. This motivates the calculation ~~of~~ the ~~average~~ relative probability increase ~~over time. The resulting factor~~ averaged over the plateau, which is estimated to ~~co~~rrespond to 25 days $\leq t \leq$ 40 days, based on Fig. S2. The resulting relative probability increase (Fig. 6) provides an estimate ~~for the question of how much~~ extent to which p-SSWs increase the probability ~~for p-NAM1000^{-}~~ of p-AO extreme events – ~~no~~t limited to a specific lag day, but time-integrated and thus independent of t. ~~If the relative probability increase is around 0, negative p-NAM1000 extremes occur after p-SSWs with a similar frequency than climatology. If the increase is larger than 0, then p-SSWs lead to a higher probability of p-NAM1000^{-} extremes.~~

~~Figure 6 summarizes this probability increase factor for different NAM1000 thresholds and for both S2S models. The estimated probability increase is computed by dividing P (as displayed in Fig. 5) by the corresponding climatological P and averaging the obtained ratios over t. In Fig. 5, the~~ Note that the measure is relative to the climatology, which also includes AO extremes that occur following SSWs. The diagnostic can therefore be interpreted as the relative probability modulation of at least one AO$^{\pm}$ event within a certain time period following the occurrence of a SSW, relative to the baseline probability where the stratospheric state is unknown.

The relative probability ~~curves were only displayed for NAM1000 < -3, whereas here, the procedure is repeated for different choices of the NAM1000 threshold.~~

—The results imply that the more negative the threshold, the greater the relative increase in probability . Even though the ~~da~~ily probability for negative p-NAM1000 values was shown to be considerably increased following p-SSWs (see Fig.4), ~~the~~ probability increase of finding increase of AO events around 0 (e.g., at least one day ~~of negative p-NAM1000 within a longer time-period is extremely high for both, climatology and following p-SSWs. Therefore the relative effect of p-SSWs is necessarily small by this measure (around +3%).In contrast, for larger negative thresholds , the effect becomes stronger as those values are generally rare and even a small influence on the distribution matters. The estimated probability increase for NAM1000 $< 2, 2.5, 3$ as revealed by UKMO is stronger compared to the ECMWF model, consistent with Fig. 4. In particular, p-SSWs increase the probability of days with NAM1000 < -3 by $\approx 40\%$ in the ECMWF and even $\approx 80\%$ in the UKMO model.~~

below/ above 0) is very small, as these events are already almost certain, even in the climatological reference. Both models show a gradual increase of relative probability of more negative AO thresholds (e.g., $\sim +35\%$ for AO< -2) and a gradual decrease for more positive AO thresholds ($\sim -40\%$ for AO$> +2$), which is consistent with a shift of the distribution toward

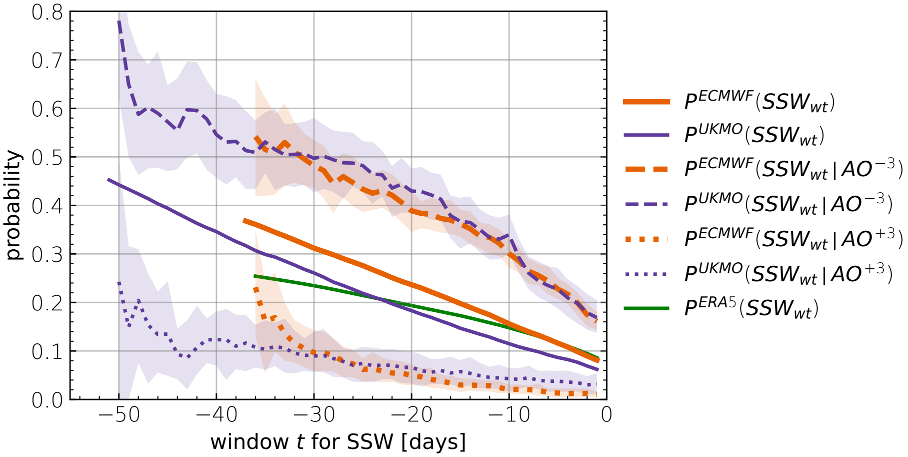

Figure 7. ~~Probability~~ Probabilities of at least one day u60 <0 within day t and day -1 relative to day ~~zero~~0, where day 0 is either a randomly sampled day ~~in ERA5 or one of the S2S models~~ (~~full lines~~solid) , an AO~~$^-$~~$^{-3}$ extreme event ~~in the S2S models~~ (dashed~~lines~~), or an AO~~$^\pm$~~$^{+3}$ extreme event ~~in the S2S models~~ (dotted~~lines~~). S2S ECMWF (orange), S2S UKMO (purple) and ERA5 (green).

more negative values. Quantitative differences in the results between the models are observed for AO thresholds of ±3. Indeed, also sampling uncertainties become considerable for thresholds greater than 2 standard deviations, as indicated by 95% confidence intervals that are obtained via bootstrap sampling among all SSW events. However, model discrepancies reach beyond the indicated confidence intervals, which will be briefly discussed in section 8.

6 ~~Predicted~~ Toward attribution of predicted AO extremes ~~preceded by predicted~~ to preceding SSWs

The last section focused on given p-SSWs and subsequent statistical signatures in AO ~~$^-$ extremes : P(subsequent AO$^-$ extreme | SSW)~~extre within a period t: $P(AO_{wt} \,|\, SSW)$. It was shown that ~~not every SSW is followed by an~~ SSWs make AO$^-$ ~~extreme (in fact, the vast majority are *not*). However, the probability relative to the baseline is increased, i.e., a SSW makes an AO$^-$ extreme~~ extremes significantly more likely.

~~A remaining open questionis: to what extent do SSWs*cause* the AO$^-$ extremes that are following them?~~

Here, we interpret "cause" in a statistical sense, which involves comparing the likelihood of occurrence of AO$^-$ extremes In this section, we will evaluate the alternative question: How many AO$^-$ events may statistically be attributed to preceding SSWs?

AO$^-$ extremes occur with and without ~~a preceding SSW. A direct cause-effect relationship between, say, a single SSW and AO$^-$ extreme is difficult to determine, although mechanistic model experiments,~~ preceding SSWs. As outlined in subsection 3.5, the distinction of whether an AO extreme was or was not exposed to a preceding stratospheric extreme requires choosing a time window for the potential exposure (e.g., ~~by nudging the stratosphere to its observed evolution have been used with some success (e.g., ??)~~was a given AO extreme preceded by a SSW within the preceding 30 days or not).

~~An important required piece of information is the baseline climatology of~~ The basis of the ~~frequency by which any random day (i.e., regardless of its AO value) is preceded by~~ evaluation in this section is that instead of conditioning on the occurrence ~~of~~ a SSW, ~~which provides an estimate of the expected chance occurrence of a SSW preceding~~ we condition on the occurrence of an AO extreme. This ~~baseline frequency is shown as a function of scanned time interval as full lines in Fig. 7 (comparing ERA5 and the S2S models). For example, the probability of SSW occurrence within 30 days preceding any random day is ≈ 0.24 in ERA5 and ≈ 0.32 in the S2S models.~~

By subtracting this baseline frequency from the probability of p-SSW occurrence preceding p-NAM1000$^-$ extremes, we may then obtain estimates of causal SSW-AO extreme relationships allows the classification of all AO events according to whether they were or were not exposed to a preceding SSW within a time window t. In total, the ECMWF analysis is based on 752 AO^{-3} and 486 AO^{+3} events, where asymmetry arises from non-zero skewness of the AO distribution (UKMO: 299 and 186).

Fig. 7 shows the probability that AO$^{\pm 3}$ events are preceded by at least one day of negative u60 within time t, corresponding to $P(SSW_{wt} \mid AO^{\pm 3})$. For example, the probability of p-SSW occurrence within 30 days preceding ~~a p-NAM1000$^-$ extreme~~ AO^{-3} extremes is close to ~~0.6 in the ECMWF S2S model (orange dashed line in Fig. 7); by subtracting the baseline estimate of 0.32 we may conclude that, based on the ECMWF S2S ensemble and the time interval of 30 days, $\approx 28\%$ of all p-NAM1000$^-$ extremes are~~ 0.5 in both models, whereas it is around 0.1 preceding AO^{+3} extremes. 95% confidence intervals, which were derived by bootstrap resampling all AO events, confirm that the diagnostics get less robust for larger time windows, due to fewer available events contributing to the AO composite. The probabilities of the extremes to be ~~*causedby a preceding p-SSW. Conversely, this implies that $\approx 72\%$ of all p-NAM1000$^-$ extremes are not caused by a preceding p-SSW (based on the 30 day time interval and the p-NAM1000$^-$ threshold)*~~ *not* preceded by at least one day of negative u60 are given by $P(\neg SSW_{wt} \mid AO^{\pm 3}) = 1 - P(SSW_{wt} \mid AO^{\pm 3})$.

~~Figure 8 summarizes the probabilities for AO$^-$ extremes that are either preceded by u60<~~ We can use the estimated probabilities $P(SSW_{wt} \mid AO^{\pm 3})$ to evaluate the fraction of attributable risk (FAR) of AO$^-$ events due to SSWs as follows[3]. First we define the FAR among the exposed[4]:

$$\mathrm{FAR}_e = \frac{\text{risk among the exposed} - \text{risk among the unexposed}}{\text{risk among the exposed}}$$

~~T~~This quantifies the fraction of AO$^-$–SSW co-occurrences ("exposed" category) that cannot be explained by internal tropospheric ~~v~~variability, where the latter risk is given by $P(AO^- \mid \neg SSW_{wt})$. An FAR$_e$ of 0 ~~by chance, caused by u60<0 and or not preceded by u60<0, as a function of the preceding period length and extended to different AOextreme thresholds. The chance occurrence~~ means that the probability of finding an AO$^-$ extreme is independent of ~~the AO threshold, but it is slightly enhanced~~

[3] In this study we neglect potential common drivers of both AO and stratospheric extremes, such as due to tropical teleconnections. Consequently our analyses of FAR may overestimate the part that is solely due to the stratosphere. Nevertheless, they serve to quantify the statistical association between stratospheric extremes and the AO, as well as quantify the predictive skill due to the stratosphere.

[4] FAR$_e$ is commonly used in climate attribution science, e.g., to determine the likelihood that an extreme weather event may be attributed to anthropogenic climate change (see, e.g., **???**).

~~in the UKMO compared to the ECMWF model. In the limit of large preceding periods, which cannot be analyzed here due to the finite maximum forecast lead time, the probability for chance occurrences is expected to saturate at 1.~~

~~Based on ECMWF forecasts, probability that~~ exposure to a preceding SSW. Likewise, an FAR$_e$ of 1 means that AO$^-$ extremes do not happen without exposure to a preceding SSW. We can estimate the involved probabilities of AO~~$^-$ extremes are caused by preceding stratospheric easterlies (u60<0)increases for larger preceding periods~~$^-$ events exposed or not to a preceding SSW using Bayes theorem:

$$P(AO^- \mid SSW_{wt}) = \frac{P(SSW_{wt} \mid AO^-) \cdot P(AO^-)}{P(SSW_{wt})}$$

$$P(AO^- \mid \neg SSW_{wt}) = \frac{P(\neg SSW_{wt} \mid AO^-) \cdot P(AO^-)}{P(\neg SSW_{wt})} = \frac{[1 - P(SSW_{wt} \mid AO^-)] \cdot P(AO^-)}{1 - P(\neg SSW_{wt})}$$

Inserting these expressions we obtain for FAR$_e$:

$$\mathrm{FAR}_e = \frac{P(AO^- \mid SSW_{wt}) - P(AO^- \mid \neg SSW_{wt})}{P(AO^- \mid SSW_{wt})} = 1 - \frac{P(SSW_{wt})}{P(\neg SSW_{wt})} \frac{P(\neg SSW_{wt} \mid AO^-)}{P(SSW_{wt} \mid AO^-)}$$

This expression involves $P(SSW_{wt})$, which represents the baseline climatology of the probability that any random day (i.e., regardless of its AO value) is preceded by a SSW within time t (full lines in Fig. 7). By definition, $P(\neg SSW_{wt}) = 1 - P(SSW_{wt})$.

Our estimates of FAR$_e$ are shown in Fig. 8a as a function of time window t, for two AO event thresholds (–2 and –3). We find that ~~the probability furthermore increases for stricter AO$^-$ thresholds, e. g. , within 28 days, 20% of~~ these estimates are not a strong function of the chosen time window. Fig. 8b summarizes the FAR$_e$ averaged over time windows of 25 to 40 days: For example, based on the ECMWF forecasts we find that on average about 50% of all AO^{-3} events that are preceded by a SSW may be statistically attributed to that SSW. For the UKMO forecasts this value is slightly higher (~60%). For AO^{-2} events these percentages are somewhat smaller but overall similar between the models. Boxplots reveal that associated sampling uncertainties are generally small, but larger for AO^{-3} events.

The attributable risk may also be evaluated for *any* AO~~<-2 events and 27% of AO<-3.5 events are caused by u60<0. Based on UKMO forecasts , the diagnostic shows a more pronounced sensitivity to the actual AOthreshold, e. g. , 14% of AO<-2 events, but 28% of AO<-3 events are caused by preceding u60<0 within 28 days.~~

Furthermore, the probabilities seem to saturate in the UKMO model for preceding period lengths $^-$ extreme (from the entire population). In this case one is interested in quantifying the fraction of AO$^-$ extremes that occur in addition to those that are "unexposed" (were not preceded by a SSW). The corresponding FAR among the population is defined as:

$$\mathrm{FAR}_p = \frac{\text{risk among the population} - \text{risk among the unexposed}}{\text{risk among the population}} = \frac{P(AO^-) - P(AO^- \mid \neg SSW_{wt})}{P(AO^-)} = 1 - \frac{P(\neg SSW \mid AO)}{P(\neg SSW)}$$

where the corresponding expressions from Bayes theorem have been inserted as before. FAR$_p$ then also quantifies the fraction of AO extremes that may be statistically attributed to a preceding SSW. For example, an FAR$_p$ of 0 means that SSWs

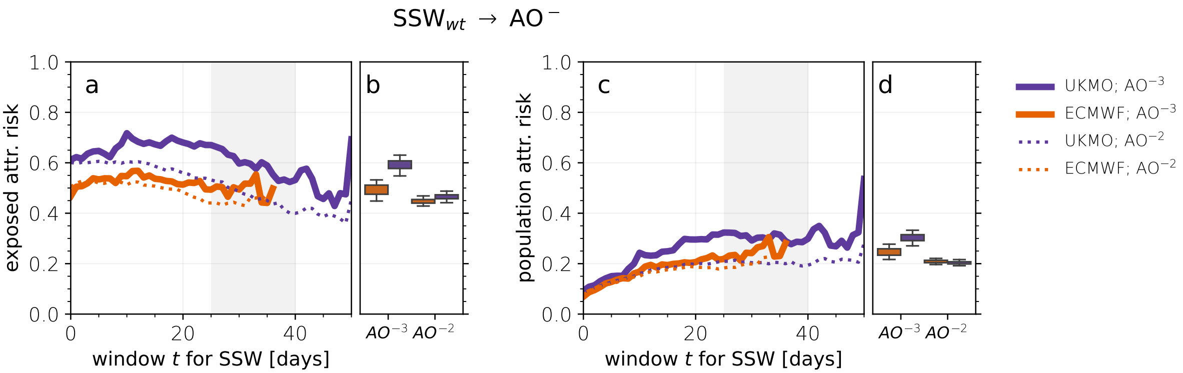

Figure 8. ~~Probability for~~ Left: Fraction of AO~~—~~$^{-2}$ (dotted) and AO^{-3} (solid) extremes ~~to be~~ that are preceded by ~~u60<0 by chance~~ a SSW within time t that may be attributed to the SSW (~~left~~ fraction of attributable risk among the exposed/ FAR$_e$, panel ~~; equal to "climatology" in Fig. 7~~a). Boxplots (quartiles 1 to 3 and 95% confidence intervals, obtained via bootstrap resampling) show FAR$_e$ averaged over time windows 25 to ~~be caused by u60~~ 40 days (~~center~~ gray shaded), as function of AO threshold (panel ~~; equal~~ b). Right: Fraction of all AO^{-2} and AO^{-3} extremes that may be attributed to ~~"~~a preceding ~~AO extreme" minus "climatology" in Fig~~SSW within time t (fraction of attributable risk among the population/ FAR$_p$, panel c). ~~7~~Boxplots (as in panel b) ~~or~~ show FAR$_p$ averaged over time windows 25 to ~~be not preceded by u60<0~~ 40 days (~~right~~ panel ~~; equal to 1 minus "preceding AO extreme" in Fig. 7~~d). Note that for larger t, ~~as a function of~~ fewer events contribute to the ~~preceding~~ diagnostics, hence, observed fluctuations for long time ~~interval~~windows t are likely related to sampling uncertainty. UKMO (purple) and ECMWF (orange).

do not increase the probability of AO extremes, whereas an FAR$_p$ of 1 means that all AO extremes may be attributed to a preceding SSW within time t.

~~—~~Figure 8c shows our estimates of FAR$_p$ as a function of time window t, similar as for FAR$_e$. As expected, estimates of FAR$_p$ are generally lower than for FAR$_e$: the likelihood of any random AO extreme to be attributable to a SSW that may or may not have happened before the AO extreme should be much smaller than that of an AO extreme that was indeed preceded by a SSW. FAR$_p$ increases somewhat with t for small t, but tends to saturate for windows longer than about ~~30 days, which cannot be tested with the ECMWF model, due to the shorter maximum forecast lead time .~~

2 weeks. For AO^{-2} events both models saturate near 0.2, whereas for AO^{-3} events they show slightly larger FAR$_p$ of around 0.25-0.3. Overall we may therefore conclude that between 20-30% of AO$^-$ extremes may be statistically attributable to a preceding SSW (within 2-6 weeks). Fig. 8d summarizes the FAR$_p$ averaged over time windows of 25 to 40 days. Despite the lower number of contributing events for larger time windows, associated sampling uncertainties are small (e.g., 95% confidence intervals for FAR$_p$ in ECMWF for AO^{-3}: [21%;28%]).

7 ~~Predicted strong~~ Strong polar vortex events and ~~related predicted, positive~~ associated AO extremes

The previous ~~section revealed that~~ sections revealed that SSWs increase the probability of subsequent AO$^-$ extremes and that a significant fraction of AO$^-$ extremes may be ~~thought of as being caused by a preceding SSW (between $\sim 25-40\%$, depending on the threshold used). Here, we extend this analysis to the~~ attributed to preceding SSWs. In the following, we summarize an analogous analysis for the statistical relationship between strong polar vortex events (SPVs) and AO$^+$ extremes.

The composite-mean evolution of p-SPVs (Fig. 9) reveals that u60 anomalies are of opposite sign, somewhat weaker in magnitude, but otherwise qualitatively similar to p-SSWs (lag 0: ~~≈~~ ∼ +20 ms^{-1} for p-SPVs; ~~≈ −~~ ∼ −30 ms^{-1} for p-SSWs, cf. Fig. 2). ~~It is observed that both~~ Both S2S models agree very well in this respect. Moreover, for negative lags, there is little difference compared to a corresponding composite based on ERA5 data, but for positive lags, u60 is slightly stronger in ERA5. The NAM response at 200hPa and 1000hPa (=AO) is qualitatively similar for p-SPVs and p-SSWs (with opposite sign), but the anomalies are again slightly weaker for p-SPVs, which is consistent with the weaker u60 anomalies (lag 21: +0.35 at 200hPa, +0.25 at 1000hPa). It is interesting that the NAM200 seems to react later to p-SPVs than to p-SSWs: While the index for p-SSWs starts to shift significantly to negative values already at lag -10 on average, a shift to positive NAM200 values for p-SPVs is observed only from lag -5 on. As with p-SSWs, the evolution of the NAM at 200hPa and 1000hPa relative to p-SPVs is less robust in ERA5 due to the smaller sample size, however, the anomalies tend to be slightly more pronounced than in the two S2S models. Overall, the composite-mean evolution of p-SPVs in the ECMWF and UKMO models appear to be consistent with real-atmosphere SPVs (as revealed by reanalysis data), as well as with previous studies (e.g., **?**).

Following the same methodology as for p-SSWs, we use the large event sample sizes to quantify the statistical relation between p-SPVs and subsequent AO$^+$ extremes. First, we quantify the relative probability increase for ~~an AO $^+$~~ at least one AO extreme after a given p-SPV within a certain time. Second, we analyze ~~whether~~ how many AO$^+$ extremes ~~are more often preceded by~~ may be attributed to preceding p-SPVs~~than any random days, in order to compute the fraction of AO$^+$ extremes that my be considered to be *caused* by p-SPVs~~.

Figure 10 shows the relative probability increase of ~~p-NAM1000$^+$ extremes after p-SPVs~~ AO extremes following SPVs relative to climatology ~~and~~ as a function of the ~~NAM1000~~ AO threshold, for both S2S models ~~: P(subsequent AO$^+$ extreme | p-SPV)~~and averaged over time windows 25 days $\lesssim t \lesssim$ 40 days:

$$\text{relative probability increase} = \frac{P(AO_{wt} \mid SPV)}{P(AO_{wt})} - 1$$

Consistent with the positive shift of the AO distribution following SPVs, the risk gradually increases for positive AO extremes, whereas it gradually decreases for negative AO extremes. For extreme thresholds of up to 2 standard deviations, the ~~probability increase of positive NAM1000 extremes after p-SPVs is similar to the probability increase of negative NAM1000 extremes after p-SSWs~~ relative probability change appears to be of similar magnitude compared to periods following SSWs (≈30-40%~~)~~, see Fig. 6). Larger thresholds reveal a reduced probability change compared to SSWs, however, 95% confidence intervals mark increasing sampling uncertainty, especially for AO$^{+3}_{wt}$ events.

~~However, for larger thresholds , the probability increase gradually diminishes again.~~

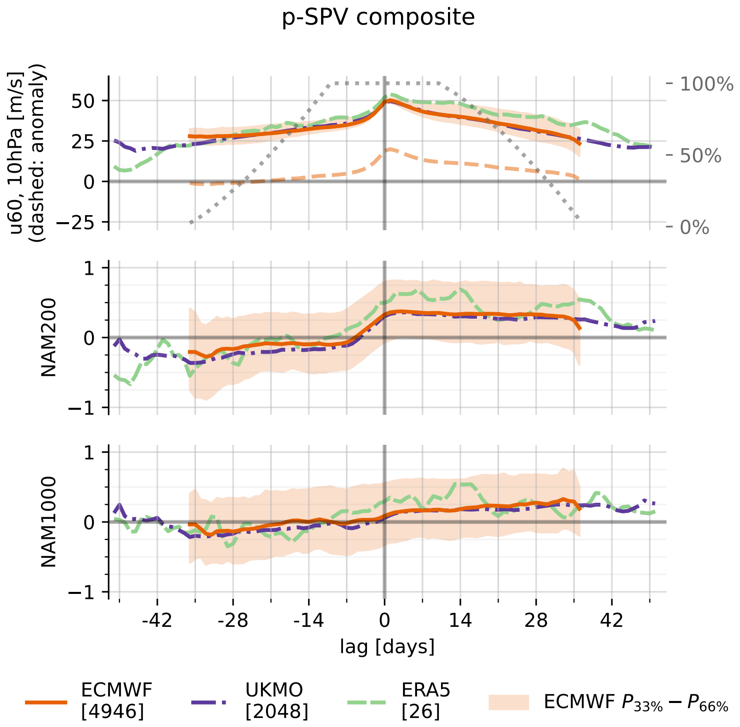

Figure 9. As in Fig. 2, for p-SPVs.

Figure 11 quantifies the fraction of positive ~~NAM1000 extreme events that are preceded by a~~ AO extremes that may be attributed to a preceding p-SPV ~~. Following the same procedure as described for negative extremes and p-SSWs~~ within a time period t:

$$\mathrm{FAR}_e = \frac{P(AO^+ \mid SPV_{wt}) - P(AO^+ \mid \neg SPV_{wt})}{P(AO^+ \mid SPV_{wt})} \tag{1}$$

$$\mathrm{FAR}_p = \frac{P(AO^+) - P(AO^+ \mid \neg SPV_{wt})}{P(AO^+)} \tag{2}$$

where FAR_e and FAR_p denote exposed and population attributable risk, as in section 6 ~~, it is observed that randomly sampled days are in about 34% of the cases preceded by a p-SPV within 28 days (left panel; UKMO: 29%), which serves as the baseline.~~

~~Ahead of positive NAM1000 extremes, p-SPV events are observed more often and the difference yields an estimate for the cases where the p-SPV is causal for the NAM1000 extreme (center panel). The results show that NAM1000 events of stricter thresholds are more often caused by a~~ for SSWs and AO$^-$ events. Among all AO^{+3} events that are preceded by at least one SPV event within four weeks, about 55% (UKMO) to 65% (ECMWF) may be attributed to the SPV (Figs. 11a, 11b). However, significant sensitivities to the exact time window are observed, as well as differences between the models. One problem is the

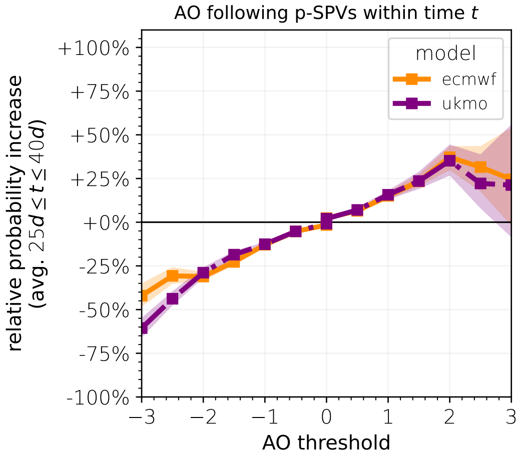

Figure 10. As in Fig. 6, for p-SPVs and subsequent ~~positive NAM1000~~ AO extremes within time *t*.

strong seasonal dependence of SPV events, as most events occur in December when the polar vortex is generally strongest. AO extremes that happen later in the winter have therefore a smaller probability to be preceded by a SPV event within a short time window than AO extremes that occur in December or January. AO^{+2} events reveal a fraction of attributable risk among the exposed to preceding ~~p-SPV. For example, about 30% of NAM1000> +2 and 46% of NAM1000> +3 events are caused by a preceding p-SPV within 28 days (UKMO: 30% and 37%). The results further suggest that the causal influence of preceding p-SPVs starts to saturate for preceding periods longer than 30 to~~ SPVs of around 40~~days. Even though the agreement between the probabilities obtained via the ECMWF and via the UKMO model is not perfect, we highlight that the estimates are still relatively close, considering that the analyses refer to the extreme tails of the pdf and only small changes therein~~% to 55%, similar to SSWs and AO^{-2} events.

Finally, ~~about 35% of NAM1000> +2 and 20% of NAM1000> +3 are not preceded by a p-SPV within 28 days. For longer periods, beyond 40 to 50 days, it appears that almost none of the positive NAM1000 events are *not* preceded by a p-SPV. This also explains the only moderate probability increase of strong NAM1000~~the fraction of all AO^{+} extremes ~~following p-SPVs in Fig. 10: In contrast to p-SSWs and subsequent NAM1000^{-} extremes , the occurrence of NAM1000^{+} extremes is in general already dominated by preceding p-SPVs. As a result, the increase in frequency is comparatively low relative to the climatology~~that may be attributed to preceding SPVs is slightly larger but similar to that for AO^{-} extremes and SSWs, with a population attributable risk of around one quarter for AO^{+2} and around one third for AO^{+3} extremes for preceding time windows of 25 to 40 days (Figs. 11c, 11d).

~~For more~~ More detailed analyses that apply the diagnostics presented in Fig. 3~~and in~~, Fig. 4 and Fig. 5 to positive AO extremes and p-SPVs ~~, the reader is referred to the~~ are shown in the supplement.

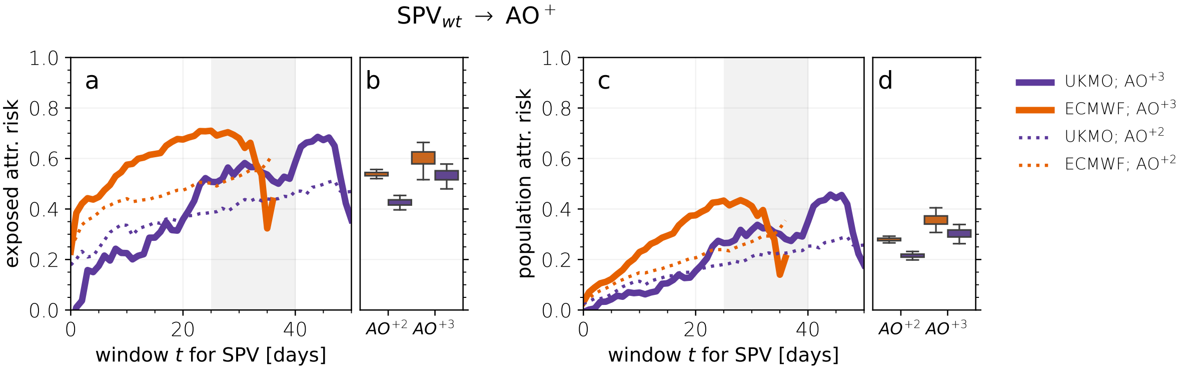

Figure 11. As in Fig. 8, for ~~u60>47 ms^{-1}~~ positive AO extremes that may be attributed to preceding ~~a p-NAM1000^{+} extreme~~SPV events within time t. ~~Analysis based on ECMWF forecasts.~~

8 Conclusions

Our results, based on a large number of extended-range ensemble forecasts, provide further evidence for stratospheric modulation of large-scale weather patterns near the surface, broadly consistent with previous results (**?**, and references therein). Previous studies generally suffer from relatively small available sample sizes, which hampers estimation of robust statistical relationships between stratospheric and tropospheric extremes (= rare events). ~~Here~~In this study, by analyzing extended-range forecast periods around predicted extreme events (e.g., p-SSWs), we effectively boost the available sample size by more than a factor of 100 and are therefore in the position to obtain robust estimates in response to our research questions:

1. By how much do stratospheric polar vortex extremes increase the probability of persistently positive or negative AO phases?

   Climatologically, 38% of negative AO phases (days with consecutive ~~NAM1000~~ AO< 0) are longer than ~~than~~ 7 days. Following p-SSWs, this is increased to 44%, which corresponds to a relative increase of 16%.

   Following p-SPVs, the probability ~~for~~ of positive AO phases that last longer than 7 days is increased from 40% to 44%.

2. By how much do stratospheric polar vortex extremes increase the probability of subsequent AO extremes?

   p-SSWs ~~are followed by AO^{-} extremessignificantly more often than expected based on climatology~~increase the probability of subsequent negative and decrease the probability of subsequent positive AO extremes. For instance, AO~~≤ -3~~$^{-3}$ events are about 40% (ECMWF forecasts) to about 80% (UKMO forecasts) more likely following p-SSWs. However, the absolute probabilities are still low, i.e., only 3.5% of SSWs are followed by AO~~≤ -3~~$^{-3}$ within four weeks, based on ECMWF forecasts (UKMO: ~~4.5~~4%).

— Following p-SPVs, the probability of ~~an AO$\geq +3$~~ AO^{+3} is increased by ~~16%~~ about 25% relative to climatology, whereas ~~—~~ AO^{-3} occur about 40% (ECMWF) to ~~55~~60% (UKMO) ~~, relative to climatology.~~

- The probability increase is smaller compared to the increase of negative AO extremes following SSWs, which is a result of the AO$^+$-extreme-climatology itself being dominated by events that follow p-SPVs.

less often.

3. ~~How often are AO extremes caused by~~ What fraction of AO extremes may be attributed to stratospheric polar vortex extremes?

About ~~one-third of AO< -3 events are, within five weeks and based on our statistical approach, caused by predicted, preceding stratospheric easterlies (u60 < 0) . Another one-third of the events is , within five weeks, not preceded by u60 < 0.~~ 50% (ECMWF) to 60% (UKMO) of AO^{-3} extremes that occur following a SSW may be attributed to that SSW (attributable risk among the exposed). 20-30% of all AO^{-3} events may be attributed to preceding SSWs (attributable risk among the population).

- ~~Within five weeks, about 45% of AO $> +3$ events are caused by a preceding SPV event .~~

~~We note that we have used the term "causality " to describe an exceedance probability relative to a climatological baseline. However, this does not rule out the existence of~~ While our stratospheric event definitions are based on absolute thresholds of the zonal-mean zonal wind, the tropospheric response is quantified via standardized anomalies of averaged geopotential. The construction of an appropriate corresponding climatology is crucial, in particular for the analysis of extreme events. However, it is also not unambiguous. Standardized anomalies are computed by normalizing differences from a population mean with the population standard deviation. The population mean as well as the population standard deviation are often a function of the season, which motivates the construction of a daily (or sometimes monthly) climatology. As the population is usually finite, any additional data point may change the population mean and will change the population standard deviation, resulting in a small adjustment of all previous (standardized) data points. On the one hand, the effect is negligible in the limit of a large population. On the other hand, it is generally larger when the additional data point is an outlier, with respect to the previous distribution. For this study, S2S forecasts were deseasonalized using the available hindcasts. The assumption is that these hindcasts sufficiently sample different kinds of variability, such that a) extreme events that occurred in individual years do not significantly distort the population distribution and thereby also the population mean and standard deviation and that b) the constructed population is robust across different initialization dates (e.g., a given event that is equally predicted at two different leadtimes corresponds to a the same standardized event in both model integrations).

Do the analyses of modulated probabilities allow conclusions about causal links between stratospheric and tropospheric circulation extremes? First, our knowledge of coupled stratosphere-troposphere dynamics suggests that a causal connection does in principle exist, although it is important to keep in mind that the coupling is mutual and causality works in both directions. Second, the concept of attributable risk allows in principle to quantify such causal links in a statistical sense, subject to filtering of common drivers. For example, ~~? have used Causal Effect Networks to analyze linear pathways that influence the midlatitude winter circulation. They find, e. g. , that~~ the ~~AO is correlated to~~ Madden-Julian Oscillation (MJO) may lead to modified risk of AO extremes (?) while at the same time modifying the likelihood of SSWs (?). On the other and, the dynamical coupling between the MJO and the AO may often involve a stratospheric pathway (?) and in such cases the stratosphere does represent a causal driver of AO modulations. Similar arguments hold for low-frequency impacts, such as Arctic sea ice concentrations (?) and the El Nino Southern Oscillation (ENSO) (?). Causal pathways may in such cases be disentangled using a causal inference-based network (?). We have carried out preliminary analyses using such a framework to distinguish causal pathways during different ENSO phases, which suggest that the ~~strength of the polar vortex and also to sea level pressure over the Ural mountains, where the latter is again correlated (with 1 month lag)with the polar vortex strength. In contrast, we focused solely on the direct statistical relationship between extreme states of the stratospheric and tropospheric circulation, with the chosen event-based approach also revealing non-linear relationships~~direct pathway *polar vortex → AO extremes* is significantly stronger than those via ENSO. A detailed analysis of these pathways is left for future work.

While the neglect of common drivers has an impact on the inferred causal links, our inferred modulated probabilities of AO extremes due to the prior occurrence of a stratospheric extreme do serve to quantify the state of the stratosphere as a predictor of subsequent AO extremes, which may be of practical value regardless of its underlying causal nature. However, even if common drivers can be neglected the statistical nature of inferred attributable risk can only quantify an *effective* causality in the following sense. Assume, for the moment, that all SSWs cause an AO$^-$ extreme, but AO$^-$ extremes additionally occur due to internal tropospheric variability. In this case some of the observed AO$^-$ extremes may have happened due to internal tropospheric variability alone while additionally be forced/enhanced by a preceding SSW. A probability analysis (e.g., estimating the FAR among the population) will then always underestimate the actual causal link and can only reveal an effective causality. This also represents a limitation of the binary classification (AO extreme / no AO extreme).

How should the observed differences between ECMWF and UKMO model be interpreted? Overall, our analyses show that the probability modulation of AO extremes up to about two standard deviations given preceding stratospheric extremes are similar between the ECMWF and the UKMO model. AO extremes of three standard deviations, i.e., $AO < -3$ and $AO > +3$ reveal discrepancies between the models. Our bootstrapping approach, e.g., for the relative probability increase (Fig. 6), shows that especially analyses based on UKMO forecasts become less robust. However, the observed discrepancies cannot be solely attributed to sampling uncertainty, given that they exist also beyond the respective 95% confidence intervals. Which model better represents the dynamics of the real atmosphere is difficult to assess, as the observational record is too short to allow for robust, similar analyses. Potential causes of the observed differences are numerous, involving differences in wave-mean flow feedbacks or external forcings, e.g., from the tropics. **?** show that the eddy kinetic energy spectrum in the ECMWF model is still in parts unrealistic and that the model may be too dissipative even at large scales, clearly indicating that models are unable to reproduce real-atmosphere dynamics perfectly accurate. **?** investigate biases in different S2S models and find, inter alia, a modest cold bias in the ECMWF and a modest warm bias in the UKMO model in the extra-tropical lower stratosphere. As the lower stratosphere has been shown to play an important role in stratosphere-troposphere coupling, we speculate that occurrences of tropospheric extremes following stratospheric circulation anomalies are sensitive to temperature biases in this region. However, a detailed analysis would be beyond the scope of this study.

In ~~this study two forecast models(ECMWF, UKMO) were considered. Given quantitative disagreements in some of the diagnostics, analyses of additional models may help to make definitive quantitative statements.~~general, we note that any two different imperfect models, will likely always reveal quantitative differences in the analysis of extreme events for a sufficiently strict extreme threshold. In the present study, we find such differences, e.g., for the relative risk, at a threshold of around three standard deviations. It is possible that more data are needed to conclusively attribute the differences to particular dynamical processes. Nevertheless, we argue that our analyses, even at a threshold of 3 standard deviations and given the associated uncertainties, are able to provide insightful quantitative estimates; especially as no obvious a priori estimate exists even for the order of magnitude of the investigated probability metrics.

~~Furthermore~~In addition to the particular points already mentioned, future work should address the question, how much of the predicted surface impact following predicted stratospheric extremes, i.e., following p-SSWs and p-SPVs, can be explained by the ~~NAM1000~~AO. Lastly, we conclude that the analysis of ~~predicted~~ *predicted* events offers potential for improved statistical characterization of other atmospheric extreme events, provided that the forecast model is capable of truthfully representing the event of interest.

Data availability. Forecasts from the S2S archive can be found at https://apps.ecmwf.int/datasets/data/s2s. ERA5 data is available at https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels.

Appendix A: Deseasonalization of S2S Forecasts

In addition to realtime forecasts, all S2S forecasting systems create also hindcasts (or "reforecasts"), which allow the construction of the respective model's climatology. In the following, we describe the procedure[5] we applied to compute a climatology of a forecast that starts on some date d (month & day of month).

1. Compute the ensemble mean of the hindcasts (Fig. A1a).

2. Compute the inter-annual mean of the hindcast ensemble means. In case of the ECMWF forecasts for example, the hindcasts cover the past 20 years (see Fig. A1b).

3. Select all (inter-annually averaged) hindcasts that start within $\pm$ 14 days relative to the date d (the start of the forecast of interest). In case of the ECMWF model, this selection subsumes 9 (inter-annually averaged) hindcasts, since hindcasts are available for every Monday and Thursday (see Fig. A1c).

4. Average the hindcasts obtained in 3, such that the forecast valid ~~time~~ times match (e.g., average forecasts for Feb 22, Feb 23, ... as opposed to matching forecast lead times, e.g., forecasts with lead time +4, +5, ..., see Fig. A1c).

[5] based on the ECMWF article "Re-forecast for medium and extended forecast range" (https://www.ecmwf.int/en/forecasts/documentation-and-support/extended-range/re-forecast-medium-and-extended-forecast-range, accessed on 23 Aug 2021).

5. Apply, to the resulting time-series, a 7-day running mean filter (Fig. A1d).

6. Due to the ±14 day window, the resulting time-series starts earlier than date d and covers a period that is longer than the forecast of interest. Cut the time-series at the beginning and at the end such that it matches the time-series of the forecast of interest. This gives the climatology (see Fig. A1d).

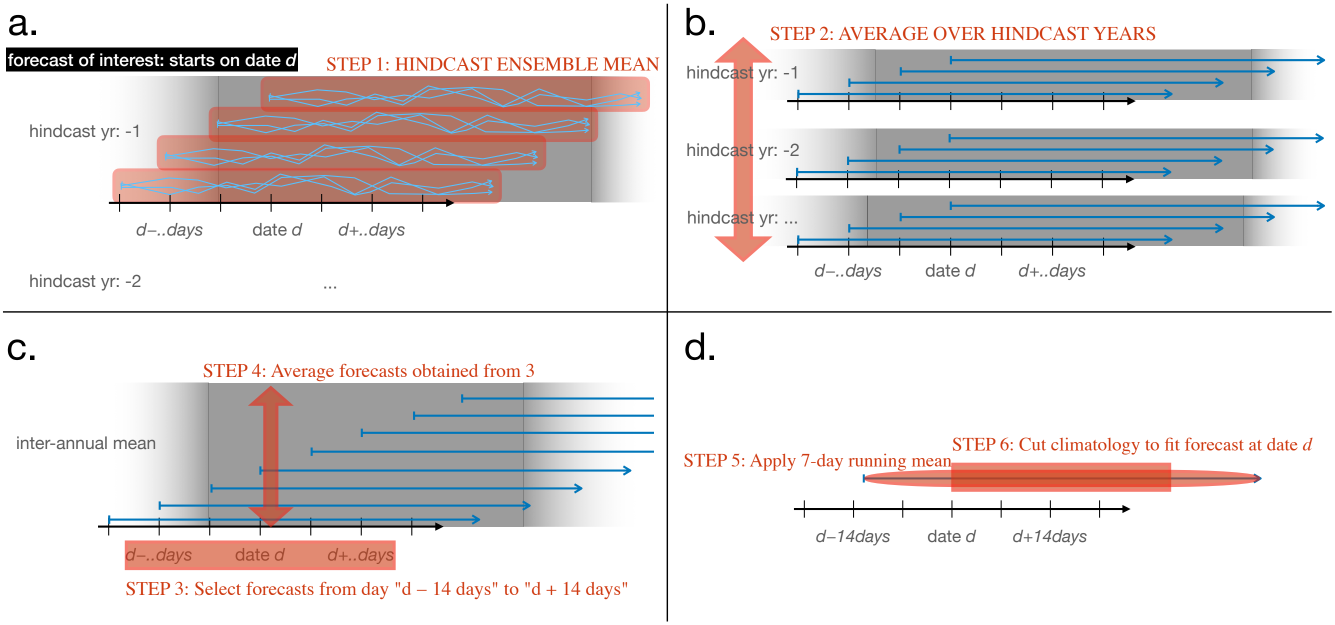
a.
forecast of interest: starts on date d
STEP 1: HINDCAST ENSEMBLE MEAN
hindcast yr: -1
d–..days
date d
d+..days
hindcast yr: -2
...
b.
STEP 2: AVERAGE OVER HINDCAST YEARS
hindcast yr: -1
hindcast yr: -2
hindcast yr: ...
d–..days
date d
d+..days
c.
STEP 4: Average forecasts obtained from 3
inter-annual mean
d–..days
date d
d+..days
STEP 3: Select forecasts from day "d – 14 days" to "d + 14 days"
d.
STEP 5: Apply 7-day running mean
STEP 6: Cut climatology to fit forecast at date d
d–14days
date d
d+14days

Figure A1. Schematic workflow for the computation of a climatology for a S2S forecast model, based on hindcasts. Gray planes illustrate that forecasts belong to the same hindcast year, where the axis from left to right denotes time and the axis from the front to the back.

Anomalies are obtained by subtracting the climatology from the raw field. Standardized anomalies can be computed by dividing the anomalies through a climatology standard deviation, ~~that is computed as the climatology~~ which is computed similar to the climatological mean, but where

- *(ad step 1)* instead of the ensemble mean, the unperturbed control run is selected (or any other single ensemble member). Using the ensemble mean would result in a too small inter-annual standard deviation at long forecast lead times (see step 2), because at long lead times, the ensemble mean *always* tends to the climatological mean state.

- *(ad step 2)* instead of the inter-annual mean, the inter-annual standard deviation is computed.

The presented deseasonalization procedure comes with several implications, for example:

- The climatologies for realtime forecasts and for hindcasts are always based only on hindcasts.

- By computing anomalies from a climatology, model errors that are a function of the season, are mitigated.

- By computing anomalies from a climatology, model errors that are a function of the forecast lead time ("model drift"), are not mitigated, because the climatology averages information that stems from different forecast lead times (see step 4).

- In case of the ECMWF model, 9 hindcast ensembles / four-week-window · 20 years · 11 ensemble member = 1980 integrations contribute to the construction of one climatology.

Appendix B: P(SSW) proxy

From observations, the annual probability of SSWs can be derived by normalizing the number of winters with SSWs with the total number of winters. In the S2S model framework, it is however less straightforward to compute the frequency of SSWs per winter, as the maximum leadtime is shorter than a winter period and many forecasts overlap. It is reasonable to tie a 0% SSW-probability to the case where there is not one ensemble member in any of the forecasts that predicts a SSW. The 100% upper boundary is less clear: Should the probability be 100% if all ensemble members in all forecasts show a SSW? In that case, a longer maximum leadtime would result in a higher SSW-probability even for the same model. Should the probability be 100% if there is at least one ensemble forecast in a winter where all members show a SSW? Again, the result would depend on the ensemble size, i.e., the technical setup, not solely on the model physics.

In this study, we compute a proxy for the model's seasonal SSW probability based on the number of SSWs per forecast day, as described in the following:

For each winter season i, forecasts with initialization dates between mid-November and mid-February are analyzed, resulting in a total of $\tilde{N} = \sum_i \tilde{N}_i$ forecast runs (counting ensemble members separately). We search for p-SSWs only in forecasts that have solely positive u60 within the first 10 days after initialization, resulting in $N = \sum_i N_i$ forecasts ($N \leq \tilde{N}$). We find E_i p-SSW events in the winter seasons, respectively, and group those by daily leadtime (similar to Fig. 1, bottom left panel), yielding $E_{i,d}$ p-SSWs in winter i at leadtime $+d$ days. As $E_{i,d}$ is approximately constant over leadtime, we compute the average number of p-SSWs in winter i per day leadtime: $E_i = \overline{E_{i,d}}$, where the overbar denotes the mean over lead times. Hence, the probability that a random forecast in winter i at a random leadtime shows a p-SSW is $p_{i,daily} = \frac{E_i}{N_i}$. The probability of no SSW for an entire winter (≈ 135 days from mid-November to end of March) is therefore $(1 - p_{i,daily})^{135}$. Finally, the probability of at least one SSW in winter i becomes: $p_i = 1 - (1 - p_{i,daily})^{135}$, as presented in Fig. 1 (top left panel). The model's average seasonal SSW probability becomes $p = [p_i]$, where the brackets denote the average over different seasons.

Note that the computed probabilities p and p_i quantify the model's tendency to predict SSWs. Particularly, this allows for inter-annual comparison and comparison between different models. However, the probabilities themselves require careful interpretation, which is why we refer to a SSW probability "proxy". Note that

- the probability quantifies SSW occurrences beyond 10 days leadtime. Thus, inter-annual variations of SSW probabilities arise only from phenomena that are predictable at more than 10 days ahead. This is also the main reason why real atmosphere SSWs have only limited effect on the computed SSW probability.

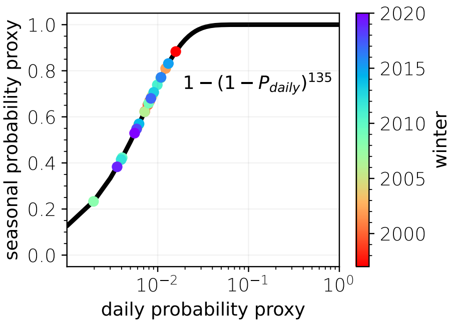

Figure B1. Estimating a seasonal SSW probability proxy based on daily SSW probabilities. Colored points show the computed seasonal probability proxy for different winter seasons as applied to the ECMWF forecasts.

- the SSW probability becomes 0% if there are no ensemble members that predict SSWs at any time beyond 10 days leadtime. A 100% probability is only reached if all ensemble members predict SSWs at each day leadtime. Fig. B1 shows the analytical relation between daily probability $p_{i,daily}$ and the associated seasonal probability p_i. For instance, a daily probability of 2% already leads to a seasonal probability of about 90%. In addition to the analytical relation, the probabilities are shown for all seasons as derived from the ECMWF forecasts.
- seasonality is not explicitly resolved in the calculations, but assumed to average out when enough forecasts are sampled.

Author contributions. JS performed the analyses under the guidance of TB. JS wrote the first draft of the manuscript. Both authors contributed to the interpretation of the results and improved the manuscript.

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

Acknowledgements. The authors thank Inna Polichtchouk for fruitful discussion on deseasonalization of S2S data. TB has been supported by the German Research Foundation (DFG) (grant no. SFB/TRR 165; Waves to Weather project). JS further appreciates the valuable scientific exchange within Waves to Weather's early career scientist program. This work is based on S2S data. S2S is a joint initiative of the World Weather Research Programme (WWRP) and the World Climate Research Programme (WCRP). The original S2S database is hosted at ECMWF as an extension of the TIGGE database. Finally, we thank the two anonymous reviewers for their constructive comments that helped to improve the manuscript.