We thank referee #2 for their review and the various suggestions to improve the manuscript. In the following we will respond in to the different comments and the general changes we intend to make to the manuscript based on them. The reviewer's comments are in black italics, our responses are shown in blue. All line numbers and references refer to the originally submitted manuscript.

This is a nice paper which investigates the potential of ensemble spread to provide useful indication of likely forecast error on S2S timescales. The methods are novel and varied, the application sound and the results should prove useful to the forecasting community. I recommend publication after considering the comments below.

We thank the referee for this encouraging assessment of our manuscript.

Main comments:

1. Fig 1 shows dominant regions of mean spread over the North Pacific and Atlantic. The shaded regions of large variability in spread lie on the flanks of these, so can the variability in spread be interpreted as (predominantly) north-south shifts of the usual regions of high spread aligning with the jets / storm tracks?

We agree that some of the patterns in Fig. 1 are consistent with meridional shifts of the storm-track-related high-spread belts in the extratropics, which enhance relative variability on their flanks. Such variability can be further amplified by remote teleconnections, for example through stratospheric downward influence. At the same time, several highlighted regions, particularly in the tropical Pacific, likely reflect more local signatures of slowly varying modes such as ENSO or the MJO. We will clarify this interpretation in the manuscript text

2. All days with lead times 14-46 days are combined in lots of the analysis here. Two thoughts on this: 1) Given the high day-to-day autocorrelation, the number of independent samples will be much less than it seems. Does this need to be taken into account anywhere? Perhaps the binning limits the impact of this. 2) The examples shown (eg fig 2a) show that the ensemble spread has saturated by day 14, which is good. This seems to be a necessary condition, as otherwise there might be a trivial link between spread and error as both are related to lead time. Can the authors confirm that saturation by day 14 is seen everywhere, not just at the couple of points shown?

We thank the referee for raising these two important points, which we address separately:

(1) Regarding autocorrelation: We agree that daily spread and error values within a given forecast are temporally autocorrelated, reducing the effective number of independent samples. However, the SRS is primarily controlled by differences between forecasts rather than day-to-day fluctuations within forecasts, as demonstrated by the close alignment between inter-forecast spread variability and the spatial structure of SRS (Section 5). Temporal autocorrelation therefore mainly affects the sampling uncertainty

of the fitted slopes rather than their expectation. To ensure that this does not materially affect the significance assessment, we performed two robustness checks. First, we recomputed SRS using only lead times beyond 4 weeks (Fig. R2.1), which reduces the number of available samples and therefore increases sampling uncertainty; despite this, the large-scale regions of significant SRS remain consistent. Second, we analysed SRS based on leadtime-averaged (weekly) spread (Fig. S3), which substantially reduces serial dependence within forecasts and yields very similar spatial patterns of significant SRS. Together, these tests indicate that temporal autocorrelation does not qualitatively alter the statistical conclusions.

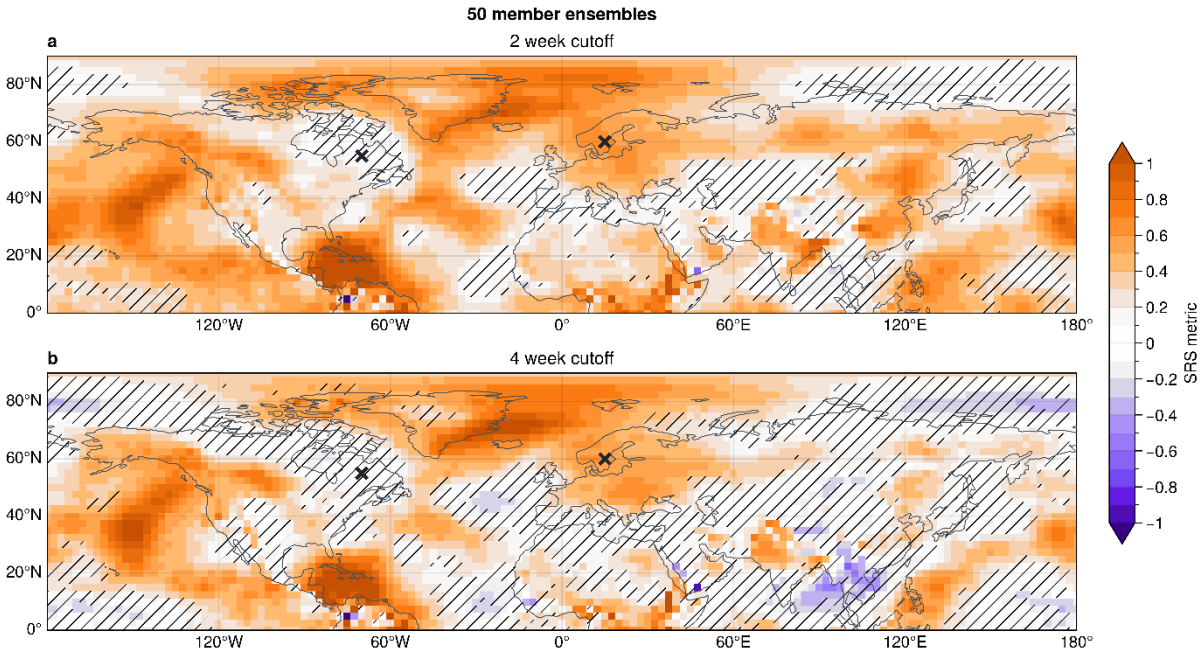

Figure R2.1: Sensitivity of the SRS metric to lead-time cutoff in 50-member ensembles. (a) SRS computed using only lead times beyond 2 weeks. (b) SRS computed using only lead times beyond 4 weeks. Hatched areas indicate regions where the SRS is not statistically different from zero at the 99% confidence level. The two maps exhibit a pattern correlation of 0.92, indicating that the large-scale spread-error relationships are robust to the exclusion of early lead times. Increased small-scale variability in panel (b) reflects the reduced number of available forecast samples.

(2) Regarding lead-time dependence: A trivial spread-error relationship could indeed arise if both spread and error were primarily controlled by lead time within the analysed window. To test this explicitly, we recomputed SRS maps using only lead times beyond 2 weeks and beyond 4 weeks (Fig. R2.1). The resulting maps show highly consistent large-scale structures, with a pattern correlation of 0.92 between the two cases. This alignment indicates that residual lead-time dependence and potential spin-up effects do not dominate the diagnosed spread-error relationships.

We note that the 4-week cutoff substantially reduces the available amount of data, leading to increased sampling variability in the SRS maps, consistent with the toy-model experiments for small sample sizes (Figs. 4g,h). Despite this increased noise, the large-scale features remain robust.

We will clarify these points the manuscript text and add robustness remarks regarding lead-time dependence and autocorrelation-effects.

3. The raw data fig 2b suggests low error values for the largest spread values (>25000), which goes against the overall relationship. Is this common or just a feature of this location?

This is in fact a slightly confusing aspect of the underlying distribution, which is not location-specific. The apparent cluster of relatively low error values at the largest spread values in Fig. 2b reflects the strong imbalance in the distribution of spread values rather than a breakdown of the spread-error relationship. Most spread-error pairs occur near the climatological spread, while extreme spread values are comparatively rare. Since the squared error for a fixed spread follows a highly skewed distribution (see pink shading in Fig. 4), this sampling imbalance can visually obscure the monotonic increase of mean error with spread in sparsely populated high-spread bins.

To test this explicitly, we performed a sensitivity experiment in which we randomly subsampled the data such that each 2500 m^2 spread interval contained at most 15 spread-error pairs (Fig. R2.2b). In this equal-count configuration, the overall increase of error with spread becomes visually clearer. The corresponding SRS values remain consistent with the full-sample estimate, although sampling variability increases substantially due to the reduced number of points and SRS can vary with different subsampling-realisations.

We will add corresponding clarifications the manuscript text.

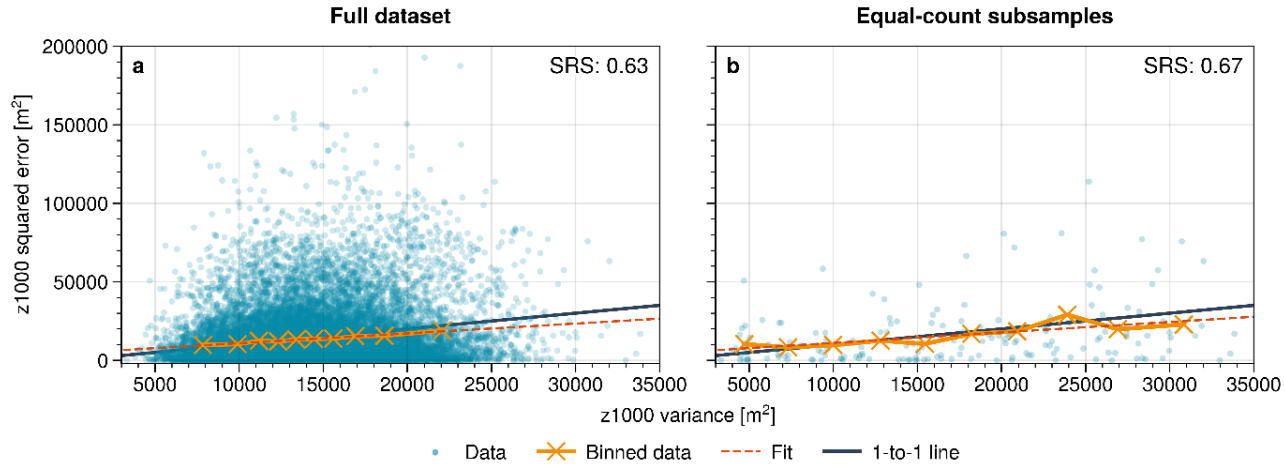

Figure R2.2: Spread–error relationship at the northern Europe example point based on (a) the full dataset and (b) random subsampling with at most 15 spread-error pairs per 2500 m^2 spread interval. Equal-count subsampling reduces the visual sampling imbalance at extreme spread values and highlights the monotonic increase of mean error with spread. The corresponding SRS values remain consistent, although sampling variability increases due to the reduced number of data points.

4. Fig 3 plots the slope for every NH point, with hatching marking points where the slope is not significantly different from zero. Does this mean that at all the non-hatched points the correlation between spread and error is significant (accounting for autocorrelation)?

The hatching in Fig. 3 indicates locations where the fitted SRS slope is not statistically distinguishable from zero at the 99% confidence level. The significance test is applied to the slope obtained from the regression of bin-mean error on spread. This is equivalent to a test of whether the Pearson correlation between bin-mean spread and bin-mean error is statistically different from zero. While this differs slightly from testing the linear relationship based directly on all raw spread-error pairs, the results are very similar in practice, and we expect the corresponding significance patterns to be largely consistent.

As in response to an earlier comment, daily spread and error values are temporally autocorrelated, which reduces the effective number of independent samples. However, we do not expect this to materially affect the large-scale significance patterns, as confirmed by robustness checks using alternative lead-time cutoffs and leadtime-averaged spread.

We will ensure the manuscript reflects these aspects.

5. I'm not sure I understand the pink shading in fig 4 - this doesn't seem to agree with the binned data shown by the black dots. In particular some of the dots with large values of spread and error seem to have very large spread values compared to the shading - is that right?

This is indeed expected behaviour and results from the highly skewed nature of the underlying distribution. Generally, the pink shading and the black dots in Fig. 4 represent different summaries of the same spread-error cloud. The pink shading shows the two-dimensional density of individual spread-error pairs and therefore highlights where samples are most frequent. The black dots, by contrast, show bin means of spread and error. These bin means can lie in regions of low density, particularly in the tails of the distribution where extreme spread values are comparatively rare. In such cases, the mean spread of a sparsely populated bin may extend beyond the visually prominent high-density region indicated by the shading. Intuitively, even in bins with large spread, most individual cases still cluster at relatively small errors (dark pink shading in Fig. 4), while a smaller number of large-error cases pull the bin-mean error upward, so that the mean values can lie well outside the highest-density region.

This behaviour is consistent with the sampling imbalance discussed in our response to the previous comment and illustrated in Fig. R2.2 for a real-case forecast, where equal-count subsampling makes the monotonic increase of mean error with spread more visually apparent.

We will clarify this in the manuscript.

6. The SRS maps in fig 6 are interesting. Eg it looks like the model spread is considerably more reliable for the northern centre of the NAO than the southern centre. Is this consistent with any other literature?

We agree that the asymmetry over the NAO region in SRS maps (e.g. Fig 3) is an interesting feature. The comparatively higher SRS over the northern centre of the NAO is consistent with the pronounced maximum in relative spread variability over the northern North Atlantic shown in Fig. 1. While enhanced variability is also present in the subtropical North Atlantic, it appears displaced further south (around 20-30°N), suggesting that different processes may be contributing there.

This behaviour is also qualitatively consistent with previous work linking subseasonal forecast uncertainty to stratospheric variability. For example, Spaeth et al. (2024) found pronounced negative spread anomalies over northern Europe following sudden stratospheric warmings, with only weak positive anomalies over the southern North Atlantic. Such results suggest that processes projecting onto NAO variability may affect the two centres of action asymmetrically. In addition, Rupp et al. (2024) showed that subseasonal spread variability over the North Atlantic is strongly linked to modulation of synoptic eddy activity, which can change the magnitude of spread-generating eddies rather than simply shifting the storm track. Such magnitude modulation provides a plausible mechanism for asymmetric behaviour between the northern and southern NAO centres.

At the same time, we caution that interpretation of detailed spatial features in the SRS maps should be done carefully, as the finite number of forecast initialisations introduces uncertainty in regional patterns.

We will add remarks to the manuscript text noting this asymmetry and a discussing possible causes.

Spaeth, J., Rupp, P., Garny, H., and Birner, T.: Stratospheric impact on subseasonal forecast uncertainty in the Northern extratropics, Communications Earth & Environment, 5, 126, 2024a.

Rupp, P., Spaeth, J., Afargan-Gerstman, H., Büeler, D., Sprenger, M., and Birner, T.: The impact of synoptic storm likelihood on European subseasonal forecast uncertainty and their modulation by the stratosphere, Weather and Climate Dynamics, 5, 1287–1298, 2024.

7. The relation to inter-over-intra variability in fig 8 is interesting. Can this be taken further back, eg to basic variances of the real atmosphere such as shown for differnent frequency bands in Blackmon et al (1984), and others?

The referee raises an interesting point. We agree that the inter-over-intra variability ratio can be interpreted in the context of classical variance decompositions of the atmosphere. Rather than adding a full frequency analysis, which would go beyond the

scope of a minor revision, we will clarify this conceptual connection in the manuscript. In particular, we intend to relate inter-variability to low-frequency geopotential height variability identified in studies such as Blackmon et al. (1984), which associate planetary-scale variability with distinct time-scale bands. This highlights that our framework effectively diagnoses where physically driven low-frequency variance modulates intrinsic predictability strongly enough to emerge above sampling noise.

8. Section 5 could be rounded off with a summary number - eg what fraction of spatial variance in SRS is explained by the perfect model test?

We thank the referee for this helpful suggestion. We will consider to add metrics that summarise some of our main key points.

9. The whole paper is framed around 'windows of opportunity', ie the low-spread end of the spectrum. Is there any interest in the high-spread end of the spectrum (walls of adversity perhaps…?)? The method uses a linear fit across the whole range of spread - do the results reflect the high-spread end of the relationship as much as the low-spread end?

This is an insightful comment and a great term for these high-spread situations. While the manuscript emphasises low-spread situations in the context of "windows of opportunity," the SRS metric is symmetric and reflects the reliability of spread fluctuations across the full range of spread values. High-spread regimes are dynamically as relevant as low-spread regimes, as they correspond to periods of enhanced intrinsic uncertainty and reduced predictability. For example, episodes of strong stratosphere-troposphere coupling during intense wave activity may lead to uncertainty in the evolution of the polar vortex and enhanced downstream tropospheric spread. We emphasise that our framework diagnoses the reliability of predictability fluctuations across the entire spectrum of atmospheric states, not solely favourable low-spread situations.

We will clarify that such high-spread states can arise from the same slowly evolving large-scale modes that create windows of opportunity, but in configurations that amplify rather than suppress error growth.

10. There are several new results given in the Conclusions & Discussion section, which are important enough to make the abstract. Consider moving these into the main paper.

We agree that the Conclusions & Discussion section (Section 6) contains new results and is comparatively long. We will revise the manuscript structure and move some of the content into the results part.

Minor:

- SRS of 0.6 is given as a summary figure in the abstract which is a nice idea but might be hard to interpret without knowing more about what SRS is.

We will revise the abstract and remove explicit SRS values.

- line 28: I would say that the whole ensemble is the 'actual prediction', not just the ensemble mean.

We agree with the referee and will revise the corresponding sentence.

- line 43: 'areas occasionally associated with anomalously low spread' are highlighted here, but could it also be occasionally high spread?

Yes, we will amended the sentence to also discuss high-spread situations.

- line 132: 'potential' windows of opportunity?

We will adept this suggestion.

- line 397-8: ref to support this statement.

We will add references to Spaeth et al. (2024a) and Rupp et al. (2024), discussing this matter.

Spaeth, J., Rupp, P., Garny, H., and Birner, T.: Stratospheric impact on subseasonal forecast uncertainty in the Northern extratropics, Communications Earth & Environment, 5, 126, 2024a.

Rupp, P., Spaeth, J., Afargan-Gerstman, H., Büeler, D., Sprenger, M., and Birner, T.: The impact of synoptic storm likelihood on Europeansubseasonal forecast uncertainty and their modulation by the stratosphere, Weather and Climate Dynamics, 5, 1287–1298, 2024.

- line 451: consider linking to https://doi.org/10.48550/arXiv.2411.17694 on signal-noise issues in subseasonal forecasts.

We thank the referee for pointing us to this relevant recent study. We will added a brief discussion linking our results to their work.

Typos:

- line 80: Ref style

Will be fixed.

- line 108: forecasts

Will be fixed.

- line 220: considerably

Will be fixed.

- line 271: intrinsic

Will be fixed.

- fig 6 caption: black line rather than grey?

Will be fixed.

- line 335: not essentially

Will be fixed.

- fig 10 caption: check line colours

Will be fixed.