Referee #1

We thank referee #1 for their review and the various suggestions to improve the manuscript. In the following we will respond in to the different comments and explain in detail the changes we made 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 in the referee's comments refer to the originally submitted manuscript, references in our responses refer to the revised version.

The manuscript "A spread-versus-error framework to reliably quantify the potential for subseasonal windows of forecast opportunity" by Rupp et al. explores the relationship between the ensemble spread and forecast error in sub-seasonal ensemble forecasts (days 14-46) by ECMWF system and in a statistical toy model. The authors propose an approach, based on spread-error relationship, to identify regions where variations in ensemble spread correlate with variations in forecast error and demonstrate, using a simple statistical model, that spread-error relationship can be deteriorated by insufficient sampling, lack of physical processes that modulate predictability, and model deficiencies.

The paper provides several interesting ideas, in particular exploring the connection between intra-forecast and inter-forecast variability of the spread, and illustrating several critical issues of sub-seasonal forecasting (such of under-sampling) using the toy model. I have no doubt that the paper should be published in WCD. However, I ask the authors to clarify several critical points before publication.

We appreciate the encouraging words of the referee and their interest in our work.

Major points:

I find that the term "the potential for windows of forecast opportunity" is obscure. I suspect that what the authors mean is "the potential to make skillful forecasts". Instead, the current message is "the potential for opportunity to make skillful forecasts". If this is really what the authors want to say, then I wonder what it means in practice.

We agree that the phrase "potential for windows of forecast opportunity" was not sufficiently explicit in the original manuscript. By this term, we do not mean forecast skill or accuracy, but the presence of substantial flow-dependent variability in intrinsic forecast uncertainty, as measured by variance of ensemble spread.

In our framework, periods of anomalously low spread correspond to windows of opportunity, while regions or situations with little spread variability cannot exhibit such windows because forecast uncertainty remains close to its climatological value at all times. The "potential for windows of opportunity" therefore refers to the variability of

spread itself, that is, the capacity of the system to occasionally enter low-uncertainty states.

We have revised the text in Sections 1, 3 and 6 to make this interpretation explicit, clarified the distinction between potential windows, realised windows and forecast skill. We also rephrased statements that could be misread as implying enhanced forecast accuracy.

The authors focus on one property of the forecast – reliability. However, I am used to think of skillful forecasts in terms of accuracy. The forecasts may lack accuracy because of low predictability even if the forecasting system is reliable. Consequently, I am used to think of windows of forecast opportunity as of periods with enhanced skill, and accuracy sufficient for decision making. I feel that your analysis, as illustrated in Figure 3, only highlights areas where physical processes modulate predictability, however it leaves open the question of whether the predictability in these regions is ever sufficient for making skillful forecasts. Therefore, I do not agree with the following statement at L468-469: "Our spread-error framework shows that, over large areas of the Northern Hemisphere, those windows are opened by slowly varying teleconnections". I feel it is difficult to discuss forecasting opportunity without analyzing accuracy (for example, anomaly correlation coefficients) and therefore I ask the authors to be more careful about their definitions and be more critical about implication of their findings.

We thank the referee for bringing up this important point and agree that forecast opportunity is often interpreted in terms of enhanced forecast accuracy. Our framework does not assess the mean or climatological level of forecast skill, but rather flow-dependent variations in forecast error and whether these variations are reliably captured by ensemble spread (spread reliability). In regions with high spread reliability, periods of reduced spread correspond to reduced forecast error, while other periods exhibit increased error. In general, we follow the established definition/interpretation of 'Windows of opportunity' used in Mariotti et al. (2020), as cited in the manuscript. We have clarified these aspects in Sections 1, 3, and 6 and revised statements that could be misread as implying uniformly high or decision-relevant forecast skill.

Mariotti, A., Baggett, C., Barnes, E.A., Becker, E., Butler, A., Collins, D.C., Dirmeyer, P.A., Ferranti, L., Johnson, N.C., Jones, J. and Kirtman, B.P., 2020. Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. Bulletin of the American Meteorological Society, 101(5), pp.E608-E625.

I am not sure why spread-error scatterplots should be made using daily values. The authors show in Figure 7 that intra-variations are spurious; thus, a lot of spread in Figure 2b is just noise. Why not define SRS using time-averaged (e.g. weekly mean) statistics?

We agree that, for subseasonal applications, time-averaged quantities such as weekly means are often more appropriate because the predictable signals of interest are typically low-frequency modes. At the same time, daily ensemble spread is routinely available and commonly inspected in forecasting practice, for example in ensemble evolution plots.

Our use of daily spread-error pairs is therefore intended as a baseline diagnostic: it allows us to assess whether the day-to-day spread fluctuations in the ensemble output contain useful information about forecast error, while also benefiting from substantially larger sample sizes. We then complement this with time-averaged analyses, which are more directly aligned with S2S predictability and, as expected, reduce spurious intra-forecast variability.

Although daily spread exhibits substantial intra-forecast variability, this variability is not uniformly dominant. Figure 7 shows that the contribution of intra-variability depends strongly on ensemble size and becomes much smaller for larger ensembles. Consistently, Figure 5 demonstrates that regions with good spread reliability can already be identified using daily values, indicating that spurious components do not necessarily obscure the spread-error relationship.

Time averaging provides an effective way to suppress spurious intra-forecast variability and can therefore enhance spread reliability. At the same time, time averaging reduces the number of independent spread-error pairs available for estimating reliability relationships, introducing a statistical trade-off between reduced noise within individual pairs and reduced sampling of the spread-error distribution. For this reason, it is not a priori clear that reliability diagrams or SRS maps based on weekly-averaged data will always be more robust than those based on daily values. We therefore consider it important to examine both daily and weekly-averaged spread, and we present time averaging as a practical improvement of spread reliability rather than as an alternative definition of the SRS.

Revised the text discussing Fig. 2 (in Section 3) and the text in (new) Section 5.5 to motivate more clearly the use of daily data in our study.

The authors make important point about time averaging (L13-14); however, this point is only illustrated by a supplementary figure (Figure S3). If the point is important enough to be elevated to the abstract, then the figure should be a part of the main manuscript.

In the revised version, we moved the discussion on weekly time averaging from the conclusions section into a separate new Section 5.5. and included former Fig. S3 as new Fig. 11.

Specific points:

L61-64: Are these assertions supported by research, or is it your hypothesis? If this is the former, a reference is needed. If this is your hypotheses, please be clear about it.

We have revised the part to be clearer about established results vs. our working hypothesis:

However, while the average ensemble spread might be a good indicator of forecast error, misrepresentations of short-term fluctuations in spread may still exist. In the following, we investigate whether such misrepresentations can limit the ability of ensembles to identify genuine windows of opportunity, since periods with low predicted uncertainty might underestimate actual errors.

L113: Provide full reference for Leutbecher et al.

We amended the bibliography entry.

L114-115: "A comparison between the IFS model and the CNRM model further shows qualitatively robust patterns (discussed in Section 6)." Robust patterns of what? Also, more information about the used CNRM data is needed.

We agree that the wording was too vague. By "qualitatively robust patterns" we refer to the large-scale spatial patterns of spread reliability (SRS) obtained from the IFS, which are reproduced in a qualitative sense by an independent CNRM ensemble dataset. We have clarified this wording in the manuscript (Sections 2 and 6) and added a brief description of the CNRM data used at the end of Section 2.

L115-116: It is quite difficult to comprehend what exactly "forecast spread reliability is influenced by the potential for windows of opportunity" means. I am not sure which definition of "reliability" the authors are using. A reliable ensemble forecast system (or any other forecast system that provides probabilistic forecasts) is one whose predicted probabilities correspond to the observed frequencies; this is what a reliability diagram illustrates. It would help if the authors provided the definition of reliability they are using. In addition, what is the difference between "windows of opportunity" and "potential for windows of opportunity"? "Opportunity" and "potential" sound synonymous to me.

We agree that the original wording was unclear and potentially confusing. The sentence referring to "forecast spread reliability being influenced by the potential for windows of opportunity" has therefore been removed.

In addition, we have clarified throughout the manuscript what we mean by reliability. In this study, reliability refers specifically to spread reliability, that is, the extent to which fluctuations in ensemble spread reliably represent, on average, fluctuations in forecast error, rather than probabilistic calibration or mean forecast skill. We now make this distinction explicit in Sections 1, 3, and 6.

We have also clarified the distinction between windows of forecast opportunity and the potential for windows of opportunity. A window of opportunity refers to a specific forecast situation with anomalously low intrinsic forecast uncertainty, while the potential for windows of opportunity refers to the variability of forecast uncertainty across different forecast situations. These definitions are now stated explicitly in the revised manuscript.

L125-127: "However, if the ensemble size is small, sampling errors will be relatively large. In such a case, some forecast/time step with, e.g., low spread, could be also associated with comparably large error, as the spread is simply underestimated due to sampling error." You assume that spread is not a good predictor for accuracy, but has this been studied? Also, how to define whether the ensemble size is small or not? The size you are using (50 members at least) does not sound small to me.

Our analysis is based on the established result that, in the limit of a perfectly reliable ensemble, the average spread equals the average error exactly (Leutbecher and Palmer, 2008). The intention of the commented passage was not to suggest that ensemble spread is generally a poor predictor of forecast accuracy. Rather, it describes a statistical sampling effect: for finite ensemble sizes, individual realisations of spread can deviate from the true uncertainty, which can obscure the spread-error relationship in individual forecasts.

We have clarified this wording in the revised manuscript and explicitly frame this as a sampling-related limitation rather than a general statement about predictability or skill. We also clarify that what constitutes a "small" ensemble size is relative and context-dependent. While 50-member ensembles are large by operational standards, sampling effects are still present, and their impact becomes more pronounced for smaller ensembles, as further illustrated by our sub-sampling experiments and toy-model results.

Leutbecher, M. and Palmer, T.N., 2008. Ensemble forecasting. Journal of computational physics, 227(7), pp.3515-3539.

Figure 2: Have you tried plotting only the "inter" component of your variance separation, rather than showing daily spread and error, which are mostly noise?

Figure 2 is intended as an illustrative example of the spread-error relationship underlying the SRS metric rather than as an optimised diagnostic. Panel (a) shows the temporal evolution of daily spread and error and therefore does not have a direct analogue for the inter component, which is time averaged by construction.

For panel (b), plotting only the inter component of spread and the correspondingly averaged error would indeed be expected to tighten the relationship and increase the SRS. However, the spread-error relationship diagnosed here is inherently statistical and holds only on average; even when using inter spread, substantial scatter between

individual spread-error pairs would remain. Using daily values in Figure 2 therefore reflects the commonly used baseline data and highlights that the relationship emerges through binning and averaging rather than at the level of individual forecasts.

The impact of suppressing intra-forecast variability, including through time averaging of spread, is examined explicitly later in the manuscript (now Figure 11, former Fig. S3), where we show that the SRS increases as expected when intra variability is reduced. We therefore chose not to add an additional inter-only version of Figure 2 to avoid redundancy. We, however, added a clarifying sentence to the discussion of Fig. 2, reinforcing that any spread-error relationship only holds in a statistical sense.

Figure 2 captions: "Red dashed line" not "Orange dashed line"

We have corrected the figure caption to avoid ambiguity in the colour description of the fitted line.

L151: How do you define "anomaly"? Figure 2 shows only positive values. For anomalies I would expect both positive (above climatology) and negative (below climatology) values.

We thank the referee for pointing this out. The displayed metrics were indeed full values, not anomalies. We have corrected the axis labels.

L175: Do you assume that ensemble mean is well represented in the toy model, or do you also assume it is well represented in operational forecasts? Is this assumption justified?

We thank the referee for critically questioning this assumption and for prompting a clarification. We now no longer rely in our argumentation on an assumption of a perfectly represented ensemble mean. Instead, we explain the behaviour of the spread–error relationship through bin averaging: sampling uncertainty affects both spread and error at the level of individual forecasts, but cancels out in bin means, while misclassification across bins systematically reduces the SRS (see response to following comment). We have revised the corresponding paragraph in Section 3 to discuss this manner and further refer to Section 6 for additional discussion.

For clarification, in the toy model we assume that ensemble members and observations are drawn from the same underlying distribution and therefore share the same population mean. The empirical ensemble mean, however, still exhibits sampling variability and is not assumed to be exact. If the forecast and observational distributions had different means, this would introduce a systematic bias and increase the forecast error, but it would not fundamentally alter the relationship between variations in spread and variations in error that underlies the SRS.

For the operational forecasts, no assumption is made about the representation of the ensemble mean, as the analysis is based entirely on the empirical relationship between ensemble spread and forecast error.

L242: Does your assumption hold? I understand that, as you under-sample the forecast distribution, the variability of the spread will in general increase. However, I believe that the variability of ensemble mean would also increase, leading to increased error. Why this would not be the case?

We have rephrased the corresponding sentence in the manuscript to no longer rely on the assumption that the ensemble-mean error distribution remains unchanged for smaller ensembles.

Instead, we now explicitly justify the behaviour of the spread-error relationship in terms of the bin-averaging procedure underlying the SRS. For finite ensemble sizes, individual forecasts can exhibit under- or overestimated spread due to sampling uncertainty, which may also be associated with larger or smaller forecast errors. However, the SRS is derived from averages over many cases with similar estimated spread. Within each variance bin, sampling errors in both spread and forecast error are random across cases and therefore largely cancel out in the bin mean. However, forecasts with over- or underestimated spread stay in the respective “wrong bin". The dominant systematic effect of reduced ensemble size is thus a reduced contrast between bins, which flattens the spread-error relationship and leads to a decrease of the SRS with decreasing ensemble size, as described in the paper. We have revised the paragraph in Section 4.2 accordingly.

To further clarify the role of sampling uncertainty of the ensemble mean and the associated correction factor, Fig. R1.1 shows an illustrative toy-model experiment. In this perfectly reliable setup, observations and ensemble members are drawn from the same distribution. Due to finite ensemble size, the empirical ensemble mean deviates from the true population mean in individual cases.

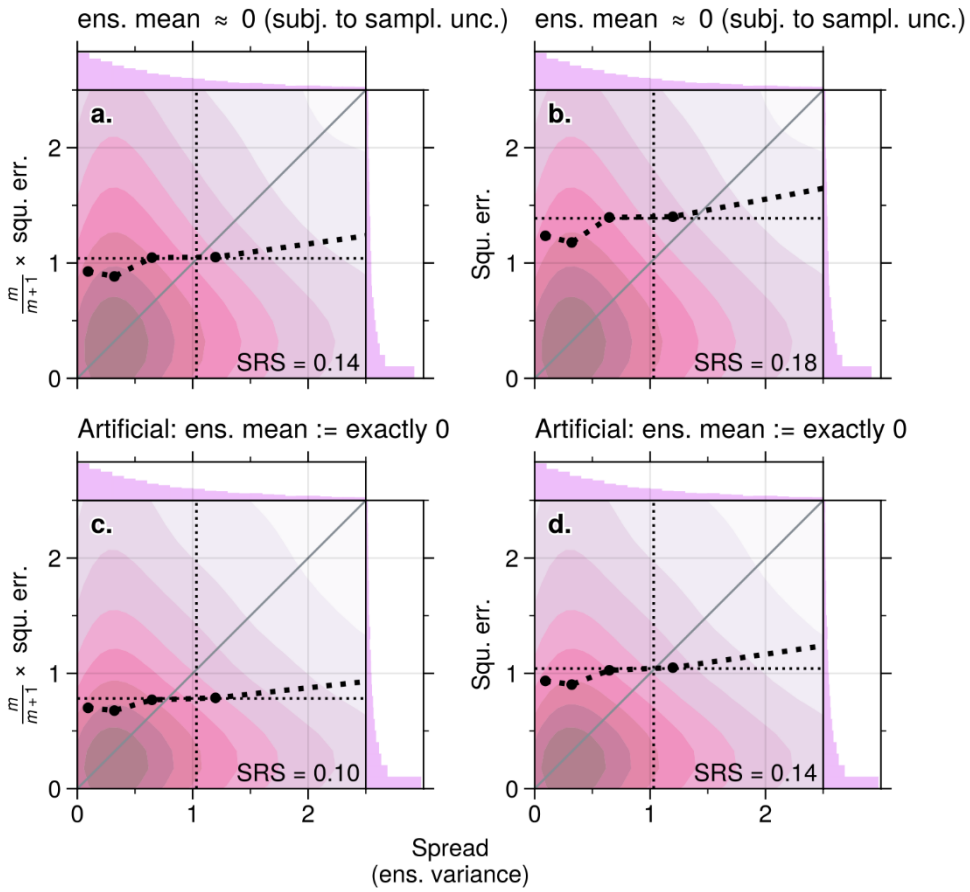

Figure R1.1: Spread-error relationship in a perfectly reliable toy-model setup illustrating sampling uncertainty of the ensemble mean due to finite ensemble size (m = 3; n = 12 000). Observations and ensemble members are drawn from the same distribution. (a,b) Squared error computed relative to the empirical ensemble mean. (c,d) Squared error computed relative to the true population mean (zero). (a,c) Error multiplied by m/(m+1). (b,d) No correction factor applied.

Figure R1.1a shows the formulation used in the manuscript: the squared error is computed relative to the empirical ensemble mean and multiplied by the factor m/(m+1) (as dicussed in the manuscript, based on Fortin et al., 2014). In this experiment we deliberately chose an extremely small ensemble size (m=3) to increase any potential effects of sampling uncertainty on the forecast error. The resulting SRS=0.14 is therefore low, but still shows limited reliability of the system.

To isolate the effect of sampling uncertainty of the ensemble mean and subsequent error, we next fix the ensemble mean to the true population mean (Fig. R1.1c), thereby removing sampling uncertainty in the mean. This modification has no systematic impact on the SRS itself, a small reduction is likely due to overall uncertainty of the system. However, because the m/(m+1) correction is still applied, the squared error is now over-corrected, leading to a downward shift of the spread-error curve relative to the 1-to-1 line. This is visible, for example, at spread = 1, where the error falls below 1.

When the correction factor is omitted in this configuration (Fig. R1.1d), the over-correction vanishes and the spread-error curve is close to the case shown in Fig. R1.1a.

For completeness, Fig. R1.1b shows the case without the correction factor but with sampling uncertainty in the ensemble mean retained. In this case, the squared error is systematically overestimated, leading to a spread-error curve shifted above the 1-to-1 line.

As stated in Section 3 of the manuscript, we consistently use this unbiased formulation, ensuring that for a reliable ensemble the expected squared error equals the ensemble variance.

Fortin, V., Abaza, M., Anctil, F. and Turcotte, R., 2014. Why should ensemble spread match the RMSE of the ensemble mean?. Journal of Hydrometeorology, 15(4), pp.1708-1713.

L251: If the error is overestimated then how this can lead to a lower error?

This sentence was indeed a bit confusing and has been revised. Forecasts misclassified as "high variance" due to sampling uncertainty will, on average, underestimate the forecast error within that "high variance bin" compared to what the spread suggests (i.e. compared to the 1:1 line). On the other hand, forecasts misclassified as "low variance" will overestimate the error compared to the 1:1 line. The combined effect leads to a systematic flattening of the spread-error relationship.

This mechanism is consistent with the behaviour illustrated in Fig. R1.1 (see discussion above), about how sampling uncertainty in the ensemble mean and error affect the spread-error relationship. We have corrected the paragraph in Section 4.2 and clarified the wording accordingly.

L235-255: I cannot understand your explanations for decreased SRS in experiment (b), and I am not sure that you can explain it without analysing variability of ensemble mean.

We have revised the description in Section 4.2 to clarify the mechanism responsible for the reduced SRS and to reduce potential confusion for the reader. In the revised text, we explicitly distinguish between sampling uncertainty at the level of individual forecasts and the bin-averaged diagnostics used to define the SRS.

We now state clearly that sampling noise affects both ensemble spread and forecast error for individual forecasts, but that these fluctuations in error largely cancel out when averages are taken over many cases within variance bins. The decrease in SRS is instead explained by sampling uncertainty in the spread estimate, which redistributes forecasts across variance bins and systematically reduces the contrast between low- and high-spread bins. This revised explanation shows that the reduced SRS in experiment (b) can be fully understood without explicitly analysing variability of the ensemble mean.

L262-270: Do you mean that a larger ensemble size than 100 members would be required to capture the spread-error relationship in the case shown in panel "c"? Have you tested this with your toy model?

We thank the referee for this question. Yes, in cases with strongly limited variability in the true variance, larger ensemble sizes are indeed required to recover the spread-error relationship. We tested this explicitly using the toy model.

Figure R1.2 shows the same setup as in Fig. 4c of the manuscript (reduced variability in true variance), but for increasing ensemble sizes. As the ensemble size increases, the SRS steadily increases and converges towards 1. This confirms that in situations with weak intrinsic spread variability, sampling noise dominates unless ensemble sizes are sufficiently large. The convergence behaviour is therefore fully consistent with the interpretation given in the manuscript.

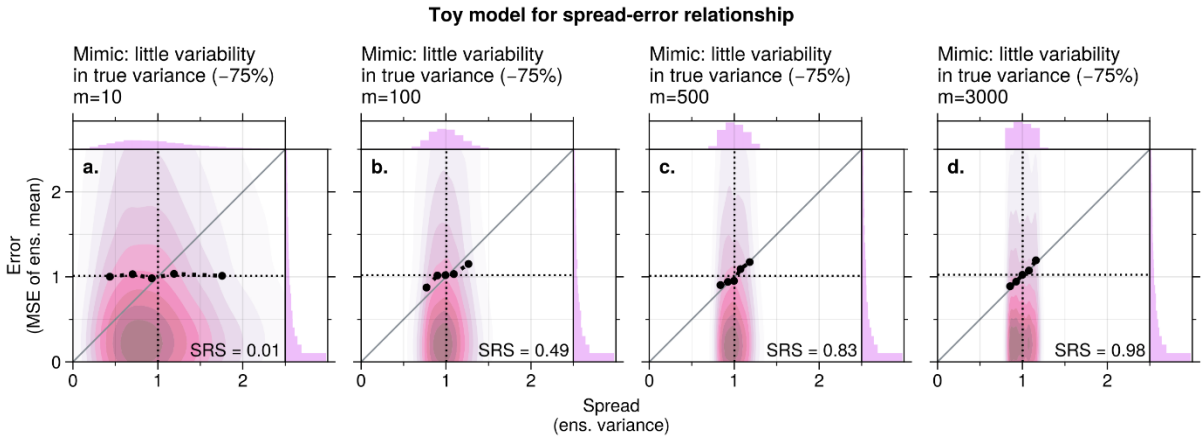

Fig. R1.2: Toy-model experiments with reduced variability in true variance (as in Fig. 4c of the manuscript) for increasing ensemble sizes (m = 10, 100, 500, 3000). The SRS increases with ensemble size and converges towards 1, demonstrating that larger ensembles are required to recover the spread–error relationship when intrinsic variability is weak.

L271: "intrincic" -> " intrinsic"

Corrected.

L289-290: Can you be more specific about which effects are unsystematic? I understand that insufficient number of cases leads to unsystematic effects, but can for example small sample size lead to unsystematic effects, or does it always lead to decreased SRS?

A smaller ensemble size always systematically decreases the SRS. We have added explicit examples to clarify what we mean by systematic and unsystematic effects. Further discussion of these mechanisms is provided in our responses to the related comments above and in the revised manuscript. Revisions for these comments will also help the reader to better understand the systematic nature of the ensemble size effect.

L324-329: Can you provide equations for the inter- and intra- variability?

We have added formal, mathematical definitions of the two variability components to Section 5.2.

L341: I do not know what the journal's policy is, but I would prefer to see the definition of the theoretical sampling error estimate in the text rather than in figure captions.

We thank the referee for spotting this. We have now added brief definition of the theoretical sampling-error scaling to the main text for clarity. The text now explicitly states that the reference sampling-error estimate is based on a 1/sqrt(M) dependence, normalised using the variability obtained for 10-member ensembles, rather than relying solely on the figure caption.

L351-352: I presume you refer to Figure 4d? It would be nice to explicitly refer to this figure in the text, for clarity.

We now explicitly refer to Figures 4c and d, rather than just the section.

L388-389: It took me a while to figure out that you are using different colour scale for Figs. 9b and 9d. I suggest using the same scale because you are making the point about smallness of the anomalies in Fig.9d, which cannot be seen with the present scales.

We agree that the use of different colour scales was not sufficiently explicit. The different scales were chosen intentionally to facilitate comparison of the spatial structures, which are very similar between the two panels, while the smaller anomaly magnitudes in the perfect-model case are evident from the reduced colour-bar range. We have clarified this choice in the figure caption.

Referee #2

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 explain in detail the changes we made 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 in the referee's comments refer to the originally submitted manuscript, references in our responses refer to the revised version.

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 now clarify this interpretation in the discussion of Fig. 1 in Section 1.

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.

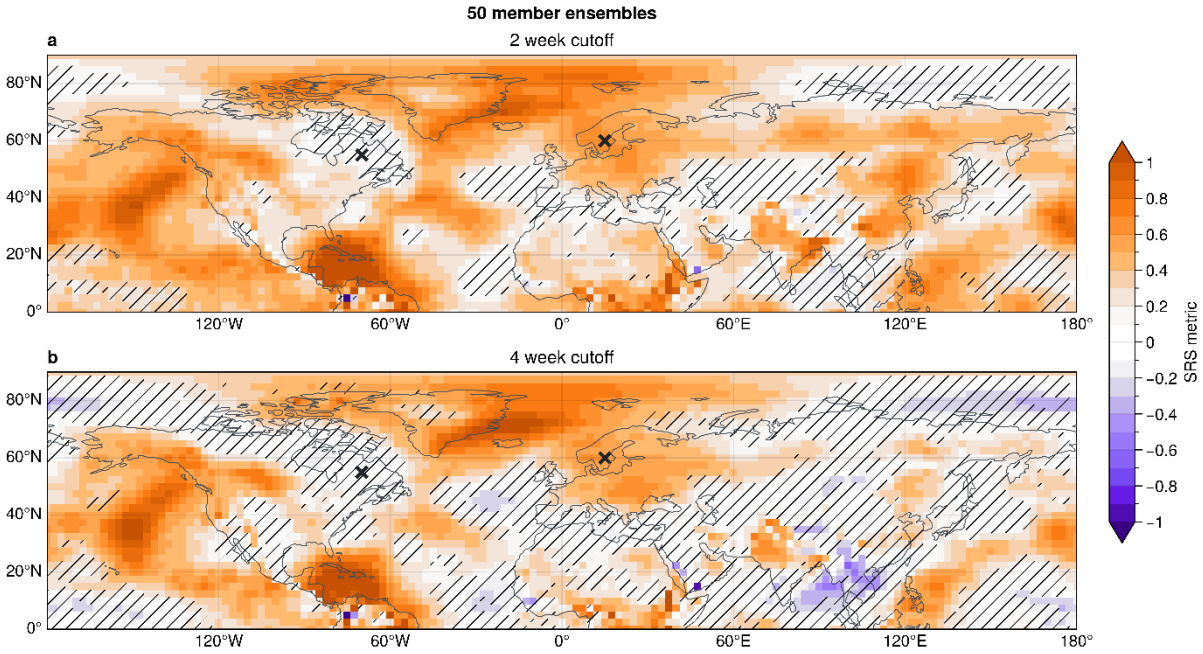

Figure R2.1: Sensitivity of the SRS metric to lead-time cutoff in 50-member ensembles. (a) SRS computed using only lead times beyond 2 weeks. (b) SRS computed using only lead times beyond 4 weeks. Hatched areas indicate regions where the SRS is not statistically different from zero at the 99% confidence 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 have clarified these points in Section 3 and added a robustness remark 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 have added a clarifying note to the discussion of Fig. 2 in Section 3.

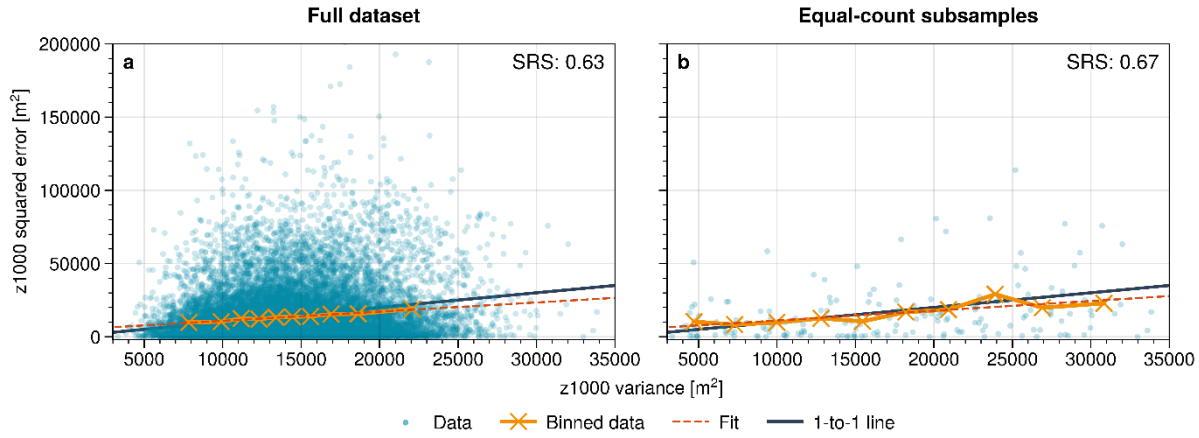

Figure R2.2: Spread–error relationship at the northern Europe example point based on (a) the full dataset and (b) random subsampling with at most 15 spread-error pairs per 2500 m^2 spread interval. Equal-count subsampling reduces the visual sampling imbalance at extreme spread values and highlights the 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 discussed in the revised Section 3 and 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 have revised the caption of Fig. 3.

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 show 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 have clarified this in the caption of Fig. 4.

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 have added a brief remark in Section 3 noting this asymmetry and a discussion paragraph interpreting possible causes and relations in Section 6.

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 have added a paragraph in Section 5.2 clarifying this conceptual connection. In particular, we now 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. To provide a compact quantitative summary, we now report the area-weighted hemispheric-mean SRS for both the reanalysis-verified and perfect-model frameworks. We find mean SRS values of 0.391 and 0.395, respectively, indicating that model-error effects modify the hemispheric-mean spread reliability by only about 1%. This suggests that intrinsic variability of the physical system largely controls the large-scale magnitude of SRS, while model errors primarily redistribute spread reliability regionally rather than substantially altering its hemispheric-mean value. We have added this clarification at the end of Section 5.3.

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. 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 have clarified in the conclusions (Section 6) 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 previous Conclusions & Discussion section (Section 6) contained new results and was comparatively long. We have therefore improved the manuscript

structure by moving some discussion point into the main analysis. In particular, we introduced two new subsections (Sections 5.4 and 5.5) that discuss (i) the proposed post-processing approach for spread calibration and (ii) the effects of temporal averaging on spread reliability. The Conclusions section has been shortened accordingly and now focuses on synthesis and broader implications rather than introducing new results. In the process, we have slightly expanded the discussion of the post-processing approach (new Section 5.4) to better describe and interpret the methodology.

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 have slightly revised the abstract and do not mention the figure any longer.

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

We agree with the referee and have revised 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 have amended the sentence. We further now discuss high-spread situations explicitly in Section 6.

- line 132: 'potential' windows of opportunity?

We adapted this suggestion.

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

We have added 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 have added a brief discussion linking our results to their work in Section 6.

Typos:

- line 80: Ref style

Fixed.

- line 108: forecasts

Fixed.

- line 220: considerably

Fixed.

- line 271: intrinsic

Fixed.

- fig 6 caption: black line rather than grey?

Fixed.

- line 335: not essentially

Fixed.

- fig 10 caption: check line colours

Fixed.