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 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.

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 will revise the manuscript text (particularly introduction and discussion) to make this interpretation explicit and clarify the distinction between potential windows, realised

windows and forecast skill. We will also make sure to rephrase 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 will clarify these aspects in the manuscript and revise 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 occasionally available and inspected in forecasting practice or operational products, 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, long-term 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.

We will revise the manuscript text discussing relevant aspects and figures (e.g. Fig. 2) 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.

We will revise the structure of the manuscript and consider moving the discussion on weekly time averaging from the conclusions section into a result section.

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 will revise this part to be clearer about established results vs. our working hypothesis.

L113: Provide full reference for Leutbecher et al.

We will amend 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 will clarify the wording in the and add a additional description of the CNRM data used.

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 wording here is unclear and potentially confusing. We will remove this sentence.

We intend to clarify 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 will also clarify 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.

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 will clarify the wording and explicitly frame this as a sampling-related limitation rather than a general statement about predictability or skill. We will 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.

We will further clarify in the manuscript text that any spread-error relationship only holds in a statistical sense.

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

We will revise the figure caption.

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 are indeed full values, not anomalies. We will correct 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. Despite stating otherwise, we do not actually rely on an assumption of a perfectly represented ensemble mean. Instead, the argument can be formulated fully via the asymmetric behaviour of sampling uncertainty in spread and forecast error during 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 will revise the corresponding manuscript sections accordingly.

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 (i.e. sample) 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 will rephrase the corresponding text in the manuscript to remove any assumption that the ensemble-mean error distribution remains unchanged despite sampling uncertainty in forecast error.

We further intend to explain in detail 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.

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.

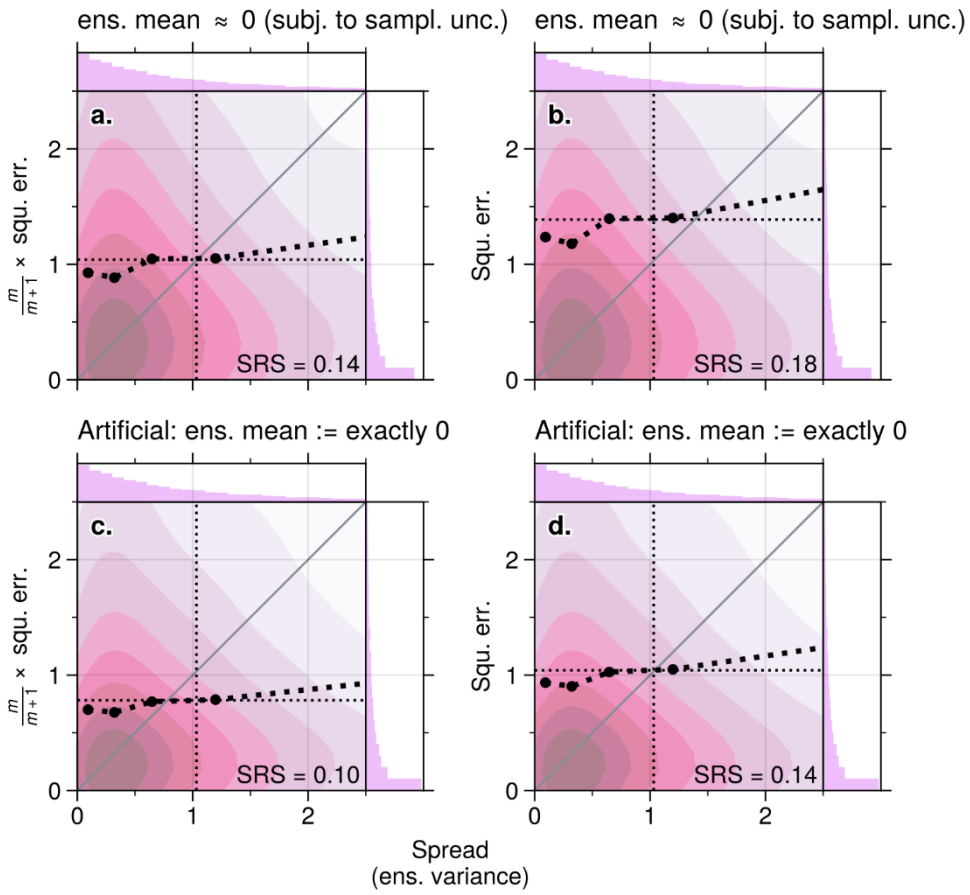

Figure R1.1: Spread-error relationship in a perfectly reliable toy-model setup illustrating sampling uncertainty of the ensemble mean due to finite ensemble size (m = 3; n = 12 000). Observations and ensemble members are drawn from the same distribution. (a,b) Squared error computed relative to the empirical 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 corresponding 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 is indeed a bit confusing and will be 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 will correct the corresponding paragraph and clarify 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 will revise the manuscript text to clarify the mechanism responsible for the reduced SRS and to minimise potential confusion for the reader. We will explicitly distinguish between sampling uncertainty at the level of individual forecasts and the bin-averaged diagnostics used to define the SRS.

We will particularly clarify 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.

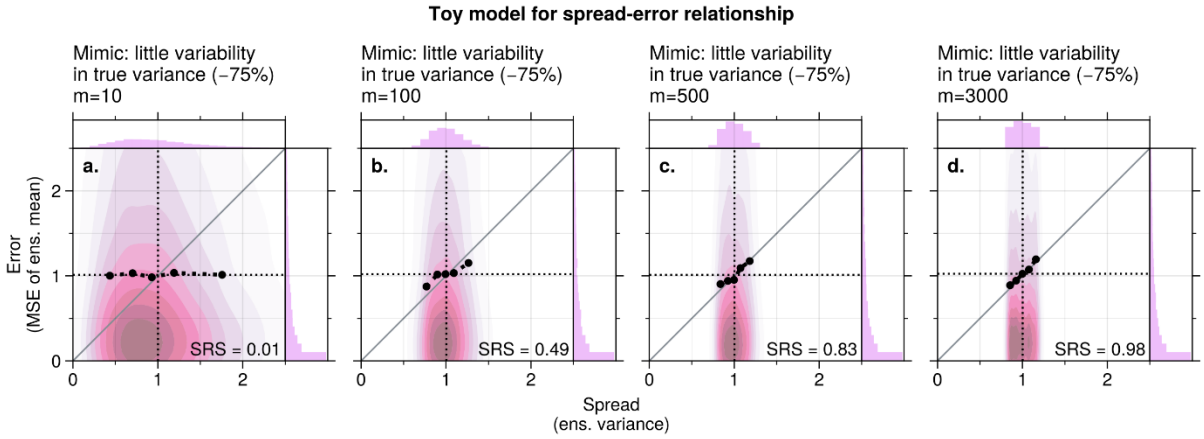

Fig. 1.2: Toy-model experiments with reduced variability in true variance (as in Fig. 4c of the manuscript) for increasing ensemble sizes (m = 10, 100, 500, 3000). The SRS increases with ensemble size and converges towards 1, demonstrating that larger ensembles are required to recover the spread–error relationship when intrinsic variability is weak.

L271: "intrincic" -> "intrinsic"

Will be 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.

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

We will add formal, mathematical definitions of the two variability components.

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 will add a brief definition of the theoretical sampling-error scaling to explicitly state that the reference sampling-error estimate is based on a 1/sqrt(M) dependence (M being the number of ensemble members), 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 will add explicit figure references.

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 will clarify this choice in the figure caption.