https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

A spread-versus-error framework to reliably quantify the potential for subseasonal windows of forecast opportunity

Philip Rupp[1], Jonas Spaeth[1,2], and Thomas Birner[1,3]

[1]Meteorological Institute Munich, Ludwig-Maximilians University (LMU), Munich, Germany
[2]Research Department, European Centre for Medium-Range Weather Forecasts (ECMWF), Bonn, Germany
[3]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany

Correspondence: Thomas Birner (thomas.birner@lmu.de)

Abstract. Mid-latitude forecast skill at subseasonal timescales often depends on 'windows of opportunity' that may be opened by slowly varying modes such as ENSO, the MJO or stratospheric variability. Most previous work has focused on the predictability of ensemble-mean states, with less attention paid to the reliability of such forecasts and how it relates to ensemble spread, which directly reflects intrinsic forecast uncertainty. Here, we introduce a spread-versus-error framework based on the Spread-Reliability Slope (SRS) to quantify whether fluctuations in ensemble spread provide reliable information about variations in forecast error. Using ECMWF S2S forecasts and ERA5 reanalysis data, aided by idealised toy-model experiments, we show that reliability is controlled by at least three intertwined factors: sampling error, the magnitude of physically driven spread variability and model fidelity in representing that variability. Regions such as northern Europe, the mid-east Pacific, and the tropical west Pacific exhibit robustly high SRS values (≈ 0.6 or greater for 50-member ensembles), consistent with robust modulation by slowly varying teleconnections. In contrast, areas like eastern Canada show little or no reliability, even for 100-member ensembles, reflecting limited low-frequency modulation of forecast uncertainty. We further demonstrate two practical implications: (i) a simple variance rescaling yields a post-processed 'corrected spread' that enforces reliability and may help to bridge ensemble output with user needs; and (ii) time averaging effectively boosts ensemble size, allowing even 10-member ensembles to achieve reliability of spread fluctuations comparable to larger ensembles. Finally, we discuss possible links to the signal-to-noise paradox and emphasize that adequate representation of ensemble spread variability is crucial for exploiting subseasonal windows of opportunity.

Keywords. forecast uncertainty, ensemble predictions, subseasonal time scales, windows of opportunity

1 Introduction

Atmospheric predictability at subseasonal timescales (about 2 weeks to 2 months) varies strongly with region and season. During certain periods, and over particular regions, intrinsic forecast uncertainty can be anomalously low, enabling higher forecast skill at these longer leadtimes (Mariotti et al., 2020). Such periods with reduced uncertainty are often referred to as 'windows of forecast opportunity'. These windows typically arise due to the influence of slowly varying atmospheric modes (Vitart and Robertson, 2018) that can be considered to exert a quasi-external influence on the region of interest. For northern hemispheric

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

mid-latitude forecasts these slowly varying modes include those of tropical origin such as the El Nino-Southern Oscillation (ENSO; Johnson et al., 2014) and the Madden-Julian Oscillation (MJO Baggett et al., 2017), and of upper atmospheric origin such as stratospheric polar vortex variability (Baldwin et al., 2003). Most studies investigating windows of opportunity analyse changes in the skill of the ensemble-mean forecast (e.g. Vitart and Robertson, 2018; Robertson et al., 2020). However, probabilistic forecasts aim to accurately capture not only the mean of the probability distribution (the actual prediction), but also higher moments such as the ensemble spread (the prediction uncertainty). Quantitative analyses of this intrinsic forecast uncertainty, as measured explicitly by the ensemble spread or variance of a forecast, are comparatively rare, and our study attempts to help fill this gap.

Previous work established that this forecast uncertainty can be highly flow-dependent (Spaeth et al., 2024b). Hence, mid-latitude tropospheric forecast spread may indeed be substantially modulated by slowly varying tropical or stratospheric teleconnections. Specifically, Spaeth et al. (2024a) recently demonstrated that the northern European subseasonal forecast uncertainty of 1000 hPa geopotential height (z1000), as directly reflected in ensemble spread, is reduced after breakdowns of the stratospheric polar vortex due to equatorward shifts of the North Atlantic storm track. Such slowly varying teleconnections and associated intrinsic predictability changes can have significant non-local impacts on ensemble spread as well. For instance, Spaeth et al. (2024a) also showed how ENSO can modulate ensemble spread in the northern Pacific. Identifying and quantifying the potential of the atmosphere to exhibit windows of opportunity might therefore be beneficial for both improving forecasts and gaining insights into the underlying dynamics of the system.

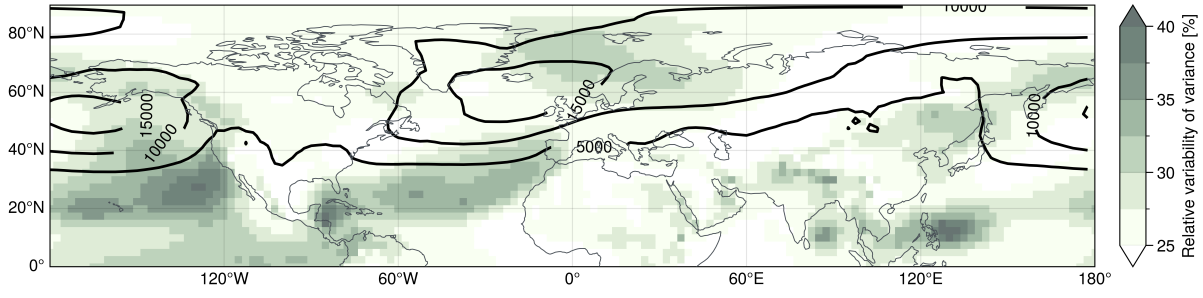

Figure 1. Intrinsic variability of subseasonal z1000 forecasts in the northern hemisphere for 50-member ensembles initialized in DJF (see Section 2 for information on the dataset). Contour lines show DJF climatological mean spread [in m^2] computed from instantaneous daily ensemble variance values averaged over subseasonal leadtimes (14 to 46 days). The shading shows relative variability in subseasonal spread, measured as standard deviation of the ensemble variance across all DJF forecast initializations and subseasonal leadtimes, normalised by the climatological average variance. Areas with large relative spread variability suggest that spread computed from individual forecasts provide added value over a fixed spread climatology, potentially identifying windows of opportunity.

Figure 1 illustrates the climatology of ensemble spread in subseasonal winter-time forecasts of z1000 over the northern hemisphere. In addition to the spread itself, it highlights regions exhibiting substantial variability in spread (i.e., areas occasionally associated with anomalously low spread) thus indicating a general potential for windows of opportunity. Prominent regions of high relative spread variability include the northern Euro-Atlantic sector, the subtropical Atlantic, and the tropical/subtrop-

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

ical Pacific. Generally, these areas are known to couple to slowly varying modes such as ENSO, MJO, and stratospheric or ocean-atmosphere variability. Some of these slow modes have been explicitly linked to enhanced forecast skill at subseasonal to seasonal timescales (e.g., Johnson et al., 2014; Baggett et al., 2017). Given this significant spatial and temporal variability in ensemble spread, an important question arises: Are these variations in ensemble spread within a model forecast reliable indicators of the uncertainty within the physical system?

Ensemble forecasting is a widely used approach to quantify forecast uncertainty and exploit windows of opportunity. The fundamental assumption of ensemble forecasting is that the spread of the ensemble (i.e., the variability across members) reflects the intrinsic forecast uncertainty or, in other words, the expected error of the ensemble mean. In an ideal, infinitely large, and statistically reliable ensemble system, the ensemble's variance across members should, on average, equal the squared error of the ensemble-mean forecast (Fortin et al., 2014). This equality is generally expected if the 'truth' is statistically indistinguishable from any individual ensemble member. If this is the case, we would consider the spread to be a reliable estimate of the forecast error.

However, real-world ensemble systems rarely achieve perfect reliability. While comprehensive diagnostics of spread-reliability are limited at subseasonal lead times, ensemble spread has been shown to lack reliability at other lead times. At short to medium lead times (up to 2 weeks), model forecasts typically show under-dispersion, where the error exceeds the error predicted by the ensemble spread (e.g. Lakatos et al., 2023). At seasonal to decadal lead times, on the other hand, atmospheric ensemble forecasts seem to be, on average, relatively reliable in their prediction of uncertainty (Weisheimer et al., 2019). However, while the average ensemble spread might be a good indicator of forecast error, misrepresentation of fluctuations of spread might still exist. Such short-term misrepresentations of ensemble spread can then limit the ability of ensembles to identify genuine windows of opportunity, since periods with low predicted uncertainty might underestimate actual errors. Ensuring that variability in ensemble spread realistically reflects variability in forecast error is therefore critical for reliable decision support.

Numerous studies have examined how the spread-error relationship varies regionally and among forecast variables, especially during Northern Hemisphere winter (DJF). Results indicate that regional variations in ensemble reliability are linked to the intrinsic variability of the ensemble spread itself. For example, Scherrer et al. (2004) analysed different uncertainty metrics for forecasts over Europe, finding generally reliable spread-error relationships. However, most previous analyses, including that of Scherrer et al. (2004), have been restricted to medium-range forecasts (up to approximately two weeks). At these shorter leadtimes, forecast dynamics strongly depend on specific synoptic conditions, with ensemble spread variations largely reflecting the error growth dynamics initiated by initial condition perturbations (Selz, 2019; Selz et al., 2022), rather than mostly describing the intrinsic predictability of the underlying physical system, which is the focus of this study.

Reliability diagrams have also been widely used in other contexts. One common approach involves comparing forecasted versus observed probabilities of specific weather events Bröcker and Smith (2007); Weisheimer and Palmer (2014). Such reliability diagrams typically assess the overall calibration of ensemble predictions but do not explicitly focus on the spatio-temporal variability of ensemble spread as done in this study. Another frequent application is the evaluation of initial perturbation schemes during short to medium-range forecasts. These analyses typically use reliability diagrams of spread and error primarily to identify under- or over-dispersion at early forecast stages, subsequently guiding adjustments to initial perturbation

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

magnitudes Leutbecher and Palmer (2008); Giggins and Gottwald (2019). In contrast, the framework applied in this study emphasises the intrinsic atmospheric uncertainty and flow-dependent variations in ensemble spread specifically at subseasonal lead times (although it can also highlight model errors in representing these processes).

The structure of the present manuscript is as follows: Sect. 2 describes the datasets and ensemble configurations. Sect. 3 introduces the spread-error reliability metric and identifies potential processes modifying reliability. Sect. 4 uses an idealized toy model to isolate if and how these processes affect reliability curves. Sect. 5 then applies the reliability diagnostics to operational subseasonal forecasts and assesses regional spread forecast skill. Finally, Sect. 6 summarizes the main findings and their implications for atmospheric dynamics and model development.

2 Data and numerical models

2.1 Re-analysis data

We use the ERA5 re-analysis dataset (Hersbach et al., 2020) of the European Centre for Medium range Weather Forecasts (ECMWF) as the representation of the atmospheric state between 1 December 2015 and 30 April 2025. In particular, we use daily snapshots of geopotential height at 1000 hPa (z1000) at 00:00UTC and daily mean 2-metre temperature (t2m) computed from 3-hourly data. All outputs are analysed on a 2.5° × 2.5° regular grid covering the entire northern hemisphere.

2.2 Subseasonal ensemble forecasts

This study uses ensemble forecasts provided by ECMWF as part of the S2S Prediction Project (Vitart et al., 2017). In particular, we use real-time forecasts initialised during boreal winter months December to February for the period from late 2015 to early 2025. The inherent horizontal model resolution is roughly 32 km, but we analyse outputs for z1000 and t2m on the same 2.5° grid as used for the re-analysis (see Section 2.1). Model output is given for 46 days after initialisation, although most of our analyses will focus on subseasonal leadtimes of 14 to 46 days.

Forecasts are initialised twice a week (Monday and Thursday) as 50-member ensembles before 27 June 2023 and daily as 100-member ensembles afterwards. Note that in this study we do not use the 'unperturbed control member' of each forecast, as it is not strictly statistically indistinguishable from the perturbed members.

To study the effect of ensemble size and forecast uncertainty we create a set of smaller ensembles by subsampling the original 100-member and 50-member ensembles. The subsampling is done by simply splitting, for example, the 100-member ensembles into two 50-member ensembles (by taking members 1 to 50 and 51 to 100, respectively). We use the same approach to subsample the 100-member and 50-member ensembles into 10 and 5 ensembles with 10 members each, respectively. The combined set of original and subsampled ensembles gives us a total of 181 forecasts with 100 members, 568 forecasts with 50 members and 2840 forecasts with 10 members. Note hereby that these sets of forecast combine different model versions, as the operationally used S2S model gets updated regularly. The subsampling approach mixes, for example, 'original' 50-member ensembles using version CY47R2 with 'subsampled' 50-member ensembles using version CY49R1. However, for the

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

purposes of this study we assume the representation of atmospheric uncertainty to vary little between different model versions, as previous studies indicate that ensemble spread characteristics remain broadly consistent across model designs. For example, Leutbecher et al. reported only small changes (a few percent) in subseasonal forecast spread when changing the stochastic perturbation scheme in the IFS model. A comparison between the IFS model and the CNRM model further shows qualitatively robust patterns (discussed in Section 6). This is partly due to the fact that forecast spread reliability is strongly influenced by the potential for windows of opportunity within the actual physical system of the atmosphere, with any model errors modifying these physical signals.

As representation of model spread in this study, we use the unbiased sample variance over ensemble members at a given time. Model errors are further given as squared error, i.e., the squared difference of the ensemble mean and the corresponding re-analysis value. The main focus of this study is then the relation of spread and squared error. Using sufficiently large sample averages, these two metrics should be equal to each other if the model accurately represents the inherent uncertainty of atmospheric evolution (as discussed in Sections 1 and 4). Note that, in contrast to some other studies, we will not analyse the relationship of spread and mean squared error given by temporal averages, but rather as averages over different atmospheric states associated with a given uncertainty. We do so by averaging over groups of forecast ensembles and/or time steps with the same ensemble variance. Forecast situations with low spread are then expected to also show, on average, low squared error. 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. This will in general weaken the expected linear relationship of spread and error and reduce the slope of spread-error curves, as further discussed in Sections 3 and 4.

3 Methodology of reliability analyses

In the present study we analyse the fluctuations in the spread of subseasonal ensemble forecasts. In particular, we present a framework to investigate the potential of the atmosphere to develop prolonged periods of reduced spread, i.e., windows of opportunity. We then use this framework to quantify the ability of the model to accurately predict such periods given limited ensemble sizes, model misrepresentation and other common sources of errors. Note that, our framework can, in principle, also be used to identify periods with anomalously high ensemble spread.

Figure 2a illustrates how the ensemble spread evolves with leadtime in subseasonal z1000 forecasts for a selected point in northern Europe. Shown are the 'climatological' spread evolution as averaged over all forecasts in DJF, as well as one specific example of the forecast initialised on 20th February 2020. For short leadtimes (up to about day 14), the climatological spread increases substantially until it converges to a roughly constant value at subseasonal leadtimes (after 14 days). The spread of the example forecast essentially follows the climatological evolution, but shows a large degree of day-to-day variability around it.

In general, the spread of an ensemble forecast should be a measure of the forecast uncertainty (cf. Section 1). We can therefore assess the reliability of predicted fluctuations in spread by comparing them to the associated forecast errors. Note that, following Fortin et al. (2014), the expected squared error of the ensemble mean equals the ensemble variance only when scaled

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

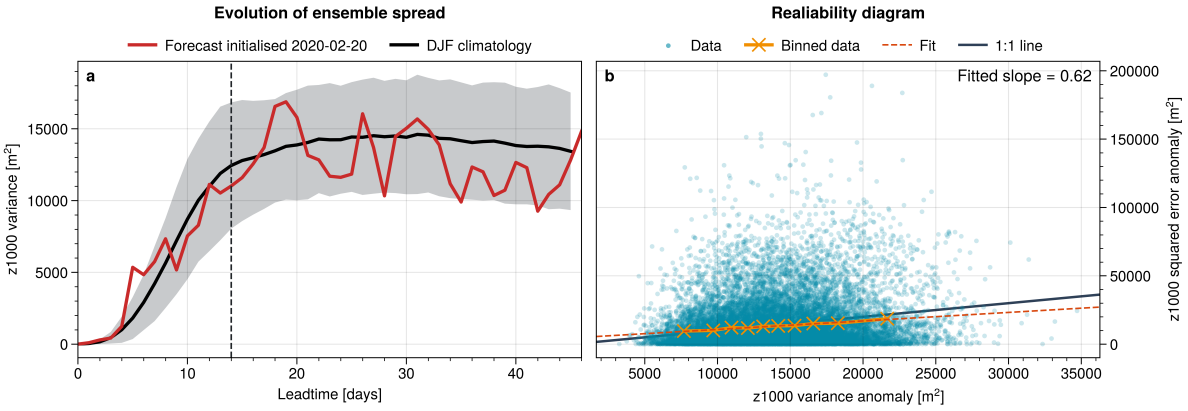

Figure 2. Visualisation of ensemble spread and spread-error relation at an example point in northern Europe (60°N/15°E) based on 50-member ensembles. (a) Evolution of spread in terms of ensemble variance as function of leadtime. Shown is the DJF climatological mean with shading indicating 1 standard deviation and an example forecast initialised on 20th February 2020. Vertical dashed line indicates 14 days of leadtime. (b) Spread-error relation with blue dots showing z1000 variance and squared error for every subseasonal leadtime (days 14-46) and all 568 ensembles in the dataset. Orange crosses show the means of 10 bins in variance direction, spaced to contain equal number of points each. Orange dashed line shows a linear fit through the bin means, with slope indicated in the top right. Black solid line shows the 1:1 line for reference.

by a factor of $M/(M+1)$ for ensemble size M. This factor essentially results from a reduction in degrees of freedom when computing the error based on an empirically estimated ensemble mean. In this study we consistently apply this "unbiasing" correction when computing the error. However, for simplicity, we will refer to this unbiased error as error (without the prefix 'unbiased') throughout the paper.

Figure 2b shows how the z1000 squared error behaves in relation to the z1000 ensemble variance for the point in northern Europe. Plotted are the squared error and spread at every subseasonal leadtime (days 14-46) within all 50-member forecasts in our dataset (see Section 2.2), we hence capture fluctuations on daily to inter-annual time scales. Since we are only interested in fluctuations of error and spread rather than climatological values, Figure 2 shows anomalies from a climatological spread/error distribution, although the methodology would work precisely the same for the full fields.

The forecast errors shown in Fig. 2 are computed based on a single observed evolution of the atmosphere and hence the resulting spread-error scatter plot shows a very disperse distribution. For a given value of ensemble variance, and under the assumption that the ensemble members are normally distributed and the model is perfect, the corresponding squared error follows a χ^2-distribution. For a reliable forecast, the mean of that error distribution should equal the associated variance value.

Indeed, when binning cases with similar spread in Figure 2, the bin-means collapse onto an almost linear relationship in good agreement with the 1:1 curve. Here we divide the distribution into 10 bins with equal number of samples in each bin. To quantify the agreement with the 1:1 line, we fit a linear function through the bin-means and extract the slope of that fit. We refer to this slope as the Spread-Reliability Slope (SRS), which represents the reliability of fluctuations in the spread of a forecast

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

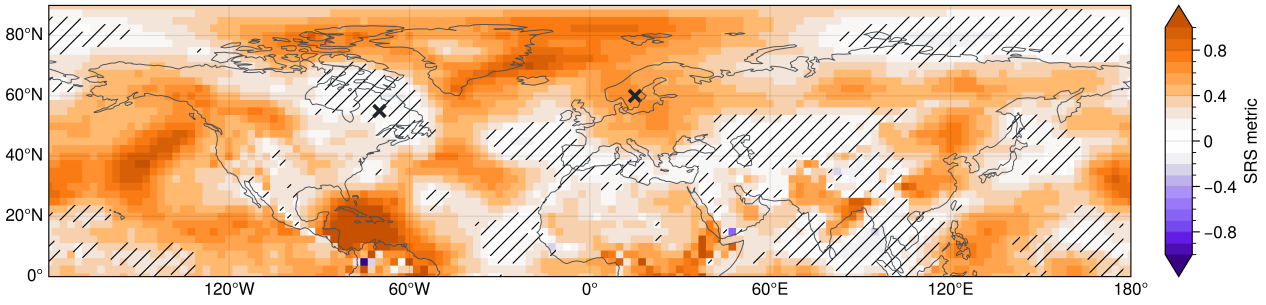

Figure 3. SRS metric computed as the slope of the linear fit to the z1000 spread-error curve at each point in the northern hemisphere. Shown are results for 50-member ensembles. Hatched areas have a slope statistically not different from 0 (at the 99% confidence level). Crosses indicate two example locations (eastern Canada at 55°N/70°W and northern Europe at 60°N/15°E) further analysed below.

at a given location. If the spread was perfectly reliable, we would have $SRS = 1$, as all bin-means would collapse onto the 1:1 line and the ensemble variance would perfectly represent the average forecast uncertainty. Any deviation of the SRS from 1 indicates a lack of reliability of the spread. In the case shown, we have $SRS = 0.62$ and hence we consider the fluctuations in spread of the forecast fairly reliable (further comparison values will be discussed below).

We can now apply the same methodology to analyse the reliability of different locations by extracting the SRS metric of the associated ensemble forecasts at these locations. Figure 3 shows that the reliability of subseasonal spread forecasts varies substantially across the northern hemisphere. For example, northern Europe has SRS values robustly exceeding 0.6, while the SRS in parts of eastern Canada is statistically not different from zero. In fact, different pronounced regions of high SRS can clearly be identified, e.g., in the northern Atlantic, the mid-east Pacific, the tropical west Pacific or the Gulf of Mexico. As discussed in Section 6, the high reliability of spread in these regions is likely given by a strong influence of slowly varying modes of variability in the atmosphere, driven by different teleconnections.

For the remainder of this study, we will investigate the different mechanisms and processes that can lead to the pronounced spatial structures in SRS seen in Figure 3. We will use various analysis methods, including toy model and perfect model approaches, to study how intrinsic variability and misrepresentation of the ensemble spread can affect the slope of the spread-error curve. Here, we assume the ensemble mean to be unbiased and a generally good representation of the predictable components of the atmosphere. Hence, all misrepresentations of the error in a forecast are assumed to be associated with unreliable spread, which would then give $SRS \neq 1$. In reality, model biases of the ensemble mean could of course play additional roles (further discussed in Sec. 6).

We identified four major mechanisms that can modify the reliability of a spread forecast and lead to $SRS \neq 1$, although other mechanisms might exist:

- **Sampling error**: random misrepresentation of ensemble spread due to small ensemble sizes
- **Natural variability**: modification of the sampling error effect due to variability of spread in the physical system

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

- **Model error**: misrepresentation of the variability of spread in the model
- **Under-representation of states**: Insufficient sampling of initial conditions producing forecasts with a given spread value

In Section 4, we start by using a statistical toy model to isolate each of these four mechanisms and study their specific effects on spread-error curves and the SRS individually. We then analyse these mechanisms and their impacts on spread reliability in operational forecast systems in Section 5. This two-step approach enables us to develop an intuitive understanding of the individual processes before quantifying their effects in more complex, real-world forecast settings, where the mechanisms typically interact and are challenging to disentangle.

4 Reliability curves studied in a toy model

4.1 Details of the toy model

In this section, we seek to develop an intuitive understanding of different mechanisms that can modify the reliability of spread in a forecast in terms of the slope of associated spread-error curves (i.e., the SRS metric). To study how individual mechanisms can affect the SRS, we use a statistical toy model that generates synthetic forecast-observation pairs with controlled properties.

We start by generating C ensemble forecast cases, divided equally into five groups. Each forecast comprises M ensemble members. The forecast value $x_{F,m,c}$ (the subscript F denotes 'forecast') for case $c \in \{1,\ldots,C\}$ and ensemble member $m \in \{1,\ldots,M\}$ is randomly sampled from a normal distribution with zero mean and a standard deviation $\sigma_{F,c}$, i.e., $x_{F,m,c} \sim \mathcal{N}(0,\sigma^2_{F,c})$. Here, we can vary $\sigma_{F,c}$ for different forecast cases, which allows us to mimic natural variability in the spread of the underlying physical system. Situations where $\sigma_{F,c}$ is small, for example, then represent periods with low intrinsic uncertainty (i.e. windows of opportunity). By varying other parameters (like M, C or $\sigma_{F,c}$) we can sample other experimental setups that mimic different characteristics of the forecasting model or underlying physical system. The values used for the different parameters are given below.

For each forecast, we then generate an observation $x_{O,c}$ (the subscript O denotes 'observation') by sampling from a normal distribution with zero mean and a standard deviation $\sigma_{O,c}$, i.e., $(x_{O,c} \sim \mathcal{N}(0,\sigma^2_{O,c}))$.

Given the ensemble members $x_{F,m,c}$ and observations $x_{O,c}$ corresponding to forecast case c, the ensemble mean is then computed as

$$\bar{x}_{F,c} = \frac{1}{M}\sum_{m=1}^{M} x_{F,m,c},$$

the ensemble variance (spread) as

$$s^2_{F,c} = \frac{1}{M-1}\sum_{m=1}^{M}\left(x_{F,m,c} - \bar{x}_{F,m}\right)^2,$$

and the squared error (SE) with respect to the observation as

$$\mathrm{SE}_c = \left(\bar{x}_{F,c} - x_{O,c}\right)^2.$$

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

Note that for a given observation-forecast pair, the ensemble variance $s^2_{F,c}$ provides an unbiased estimate of the underlying forecast variance $\sigma^2_{F,c}$. However, the accuracy of this estimate improves with increasing ensemble size.

Our modelling strategy within the toy model includes a reference experiment and different perturbation experiments, where we vary individual parameters to simulate changes to the model and physical system and isolate their effect on the resulting spread error curve. The reference case is supposed to represent an 'ideal case', associated with a spread-error slope of exactly one ($SRS = 1$). The parameter setup for this reference case is given in Table 1a.

Table 1. Overview of toy model experiment configurations, with the top row describing the reference experiment. Here, M denotes the ensemble size, C the number of forecast-observation pairs, σ_O the standard deviation of the observations, and σ_F the standard deviation of the forecast ensemble members. The two experiments mimicking few verification dates (g/h) are initialised with different random seeds, but use otherwise equal parameters.

Experiment	M	C	σ_O	σ_F
(a) Ideal spread-error relation	100	3000	$\{0.7, 0.85, 1, 1.15, 1.3\}$	as σ_O
(b) Mimic small ens. size	**10**	3000	$\{0.7, 0.85, 1, 1.15, 1.3\}$	as σ_O
(c) Mimic little variability in observed variance	100	3000	$\mathbf{\{0.925, 0.9625, 1, 1.0375, 1.075\}}$	as σ_O
(d) Mimic large variability in observed variance	100	3000	$\mathbf{\{0.475, 0.7375, 1, 1.2625, 1.525\}}$	as σ_O
(e) Mimic model error (too little variability)	100	3000	$\{0.7, 0.85, 1, 1.15, 1.3\}$	$\mathbf{\{0.925, 0.9625, 1, 1.0375, 1.075\}}$
(f) Mimic model error (too large variability)	100	3000	$\{0.7, 0.85, 1, 1.15, 1.3\}$	$\mathbf{\{0.475, 0.7375, 1, 1.2625, 1.525\}}$
(g/h) Mimic few verification dates	100	**60**	$\{0.7, 0.85, 1, 1.15, 1.3\}$	as σ_O

Here, we choose a large ensemble size of $M = 100$, which corresponds to the ensemble size of the latest operational sub-seasonal forecast system at ECMWF. Lower values of M then model smaller ensembles. We further choose a case sample size of $C = 3000$, considerable larger than the number of 100-member forecasts analysed in this study (which is 181 for 100-member ensembles; see Section 2.2). Modifying C allows us to to model the effect of this reduced sample size of cases. Within the reference experiment, we then vary the variability of observed spread $\sigma_{O,c}$ by choosing values from the set $S = \{0.7, 0.85, 1.0, 1.15, 1.3\}$. Specifically, the first $C/5$ forecasts use $\sigma_{O,c} = 0.7$, the next $C/5$ forecasts use $\sigma_{O,c} = 0.85$, and so on, with the final $C/5$ forecasts using $\sigma_{O,c} = 1.3$. Our qualitative conclusions are not sensitive to the precise distribution of the set S, but it can modify some aspects of the shape of the spread-error curves. However, we run experiments with a reduced or increased range of values in S to mimic underlying physical systems with low or large variability in their intrinsic uncertainty, respectively. These different physical systems could represent different spatial locations or periods in different seasons. Additionally, we can choose the forecast distribution (i.e., $\sigma_{F,c}$) to exactly match the observed distribution $\sigma_{O,c}$ (which simulates a perfect model), or follow different distributions (which simulates model error). A summary of the different experiments and their associated parameter combinations is given in Table 1.

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

4.2 Sensitivities of spread-error curves in the toy model

Figure 4a shows the spread-error curve for the reference toy model experiment. This reference experiment uses a large ensemble size ($M = 100$) with good sampling (C=3000) and a correct model representation of the forecast distribution ($\sigma_{O,c} = \sigma_{F,c}$). Therefore, the spread-error curve lies almost exactly on the 1:1 line, as expected, with spread-reliability-slope of $SRS = 0.99$.

If we reduce the ensemble size to 10 members (Figure 4b), the spread-error curves becomes more shallow and hence the SRS reduces as random differences between the ensemble sample variance ($s^2_{F,c}$) and the underlying forecast population variance ($\sigma^2_{F,c}$) become larger. There are two perspectives to think about how this insufficient-sampling effect reduces the slope of the spread-error curve. First, increase in variability of the ensemble sample variance stretches the distribution in x-direction, which can be intuitively understood as follows. The naturally varying uncertainty in the system would allow variances in the range $[\min_c \sigma^2_{O,c}, \max_c \sigma^2_{O,c}]$. However, in extreme cases where $s^2_{F,c} \gtrsim \min_c \sigma^2_{F,c}$ or $s^2_{F,c} \lesssim \max_c \sigma^2_{F,c}$, sampling error might add on to the modelled variance and hence the range of output variances would include $s^2_{F,c} < \min_c \sigma^2_{F,c}$ and $s^2_{F,c} > \max_c \sigma^2_{F,c}$. Since we assumed the ensemble mean to be well represented within our toy model the error distribution stays unchanged. Consequently, the slope of the spread-error curve will reduce with increasing sampling error, in other words, SRS will decrease with decreasing ensemble size.

As a second perspective on the SRS reduction, consider that in a system with perfectly reliable spread (in which $\sigma^2_{O,c} = \sigma^2_{F,c}$), the forecasts with the largest true variance, $\sigma^2_{O,c}$, should lie on the 1:1 line when plotting variance against the case-averaged squared error. However, since neither $\sigma^2_{O,c}$ nor $\sigma^2_{F,c}$ are known in practice, we use the ensemble sample variance $s^2_{F,c}$ as an estimate. While this estimate is perfect in each single case in an infinitely large ensemble, it is subject to sampling noise when the ensemble size is finite. As a result, averages over cases with the highest spread values (i.e., those with the largest ensemble spread $s^2_{F,c}$) will not only include cases with genuinely large forecast variance $\sigma^2_{F,c}$, but also cases with smaller forecast variance that appear inflated due to sampling fluctuations. These misclassified forecasts tend to overestimate their expected error, thus lowering the average error for that variance value and pulling it below the 1:1 line. The reverse happens for averages over cases with the lowest spread values: it may include forecasts with higher true variance that were underestimated due to sampling variability. These cases increase the average error in the bin, pulling it above the 1:1 line. The net effect is a systematic reduction of the SRS.

This behaviour is consistent with ensemble sampling theory, which provides insight into why forecasts with few members often misrepresent variability in spread. With small ensemble sizes M, the sample variance becomes noisy (high "variance of the variance"), and individual forecasts may, by chance, fail to sample extreme outcomes, causing the actual error to exceed the predicted spread. Increasing M reduces this sampling error, as the standard error of the spread estimate decreases proportional to $1/\sqrt{M}$ (cf. Tempest et al., 2023). In general, larger ensembles are therefore required to obtain a stable spread–error relationship, especially for higher-order moments such as the ensemble variance.

The impact of sampling noise due to finite ensemble size also depends on the intrinsic variability of the true uncertainty, $\sigma^2_{O,c}$, which is determined by the characteristics of the underlying physical system. If $\sigma^2_{O,c}$ varies strongly across cases (i.e., if the range $\left[\min_c \sigma^2_{O,c}, \max_c \sigma^2_{O,c}\right]$ is large) then the relative effect of sampling noise becomes less significant. In such cases,

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

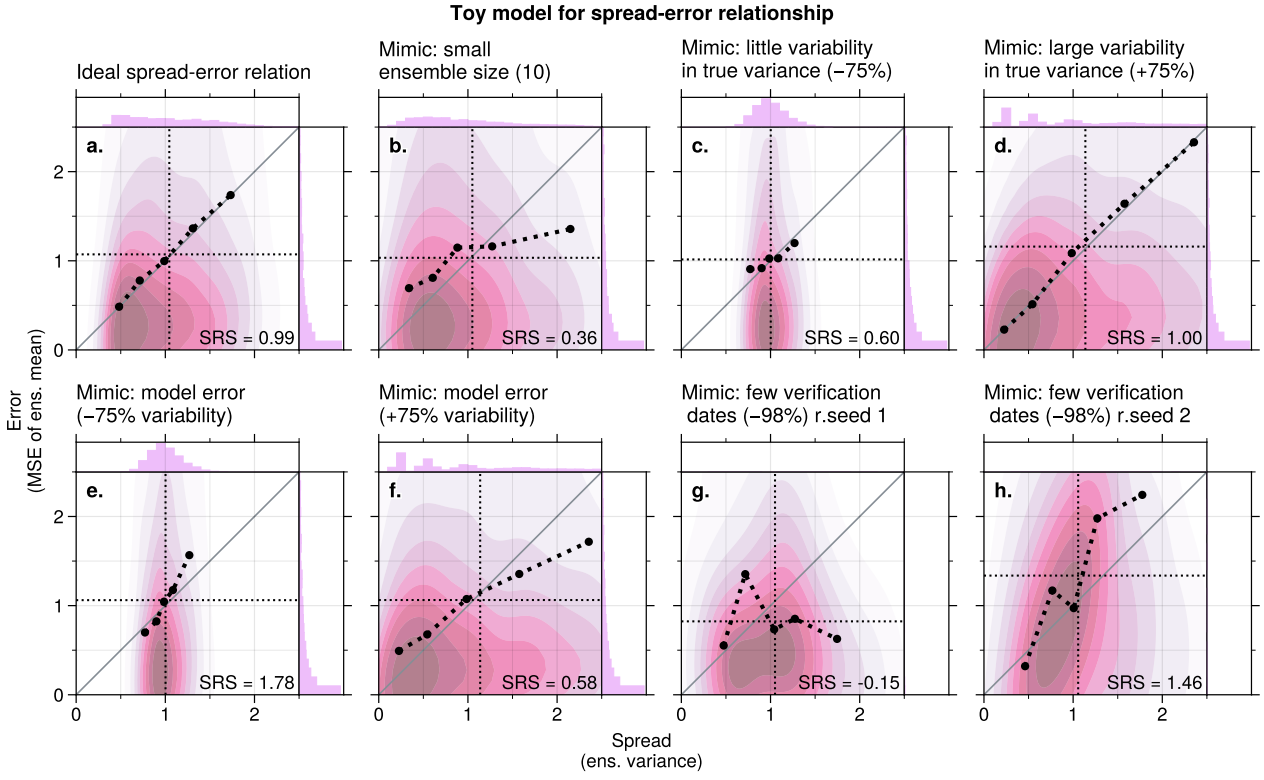

Figure 4. Toy model experiments illustrating the spread–error relationship (ensemble variance vs. mean squared error of the ensemble mean). Pink shading shows the 2D distribution of individual spread-error pairs. Spread-error pairs are grouped into five bins based on the ensemble spread; black dots indicate the average spread and error within each bin. For visual guidance, these bin means are connected with thick dashed black lines. Thin black dotted lines indicate the overall mean spread (vertical) and error (horizontal). The solid grey diagonal line represents the 1:1 relationship. The spread-reliability-slope (SRS) is indicated for each case.

the differences between $\sigma^2_{F,c}$ and its noisy estimate $s^2_{F,c}$ are small compared to the variability in $\sigma^2_{F,c}$ itself. As a result, even with a finite ensemble size, the binning of forecasts by spread is more robust, and the spread-skill relationship appears less distorted. This effect is illustrated in Figures 4c,d. In panel 4c, $\sigma_{F,c}$ varies within a narrow range (0.925 to 1.075), while in panel 4d it spans a much broader range (0.475 to 1.525). The comparison shows that larger variability in spread of the physical system improves the clarity of the spread-skill relationship under finite ensemble conditions, mitigating the SRS reduction due to sampling error.

In addition to sampling error and natural variability in spread, model biases in how intrincic uncertainty responds to physical drivers can also affect the spread-error relationship. For instance, anomalies in ensemble spread may be systematically too small if the model responds too weakly to teleconnection patterns. The opposite would be true if the model responds too strongly to teleconnections. In our toy model, such biases can be mimicked by choosing different values for the variance of observed ($\sigma_{F,c}$) and modelled ($\sigma_{O,c}$) distributions. Figures 4e and f show experiments where the model over- or underestimates the anomalies

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

in spread. Such a misrepresentation of the spread leads to a stretching or compression of the distribution in variance-direction. Analogous to the effect discussed with regards to sampling error, this will affect the slope of the spread-error curve and alter the SRS. In general, an over-estimation of spread variability (so $\sigma_{F,c}/\sigma_{O,c} > 1$ and a stretching in variance direction) will lead to $SRS < 1$, while an under-estimation of spread variability will do the opposite and lead to $SRS > 1$. Note that here we are discussing over- and under-estimation of the variations in spread, and not an overall over- or under-estimation of the mean spread (which may or may not be accurate on average).

Next we analyse the effect of a limited number of cases, i.e., few forecast-observation pairs (small C). This will in general lead to a violation of the underlying equality between errors and spread (see Section 3) and introduce random deviations of the spread-error curve from the 1:1 line. Since these deviations are unsystematic, they can randomly lead to increases or decreases of the SRS. A system with an under-representation of cases can therefore, in principle, produce an SRS larger than 1 (see Fig. 4h), smaller than 1 or even smaller than 0 (see Fig. 4g). Note that in situations with small C the SRS could take very large or very small (or even negative) values despite the underlying ensemble forecast having many members and no model error.

The toy model presented in this section allowed us to study the effects of different mechanisms on the spread-error curves and the SRS in an isolated manner. While some of these effects are systematic and always increase or decrease the SRS, others are unsystematic. A forecasting system will typically suffer from multiple error sources. This can lead to a superposition of the corresponding effects and hence SRS values that either deviate strongly from one for multiple reasons, or SRS values close to 1 despite major error sources due to cancellation. The next section analyses the reliability of spread forecasts in subseasonal ensembles by trying to disentangle the different mechanisms and studying their potential importance individually.

5 Reliability of operational forecasts

5.1 Sampling error due to ensemble size

Various mechanisms can affect the reliability of spread forecasts and lead to deviations of the SRS from unity, as shown within the toy model in Section 4. This section goes through the list of individual mechanisms and analyses their importance within subseasonal ensemble forecasts of the real atmosphere.

We start by analysing the effect of sampling error on the SRS. Figure 5 shows the reliability of z1000 spread within the northern hemisphere for three different ensemble sizes. It can be seen that the SRS generally increases with ensemble size. While 10-member ensembles show poor spread reliability almost throughout the entire hemisphere (SRS close to zero), 50-member ensembles exhibit substantially more reliable spread in various regions (e.g. northern Europe, eastern Asia, western North America or around the Gulf of Mexico), with SRS closer to 1. We see further SRS improvements in many of these regions, when increasing the ensemble size to 100 members. However, for 100-member ensembles we find regions with SRS larger 1 or negative SRS. This is likely due to the limited number of 100-member ensembles available and reflects an under-representation of atmospheric evolutions in the system (cf. Fig. 4c).

Although 50 and 100 member ensembles show generally reliable spread in many regions, other regions do not exhibit visible improvements with increasing ensemble size. A pronounced region in eastern Canada, for example, is associated with a slope

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

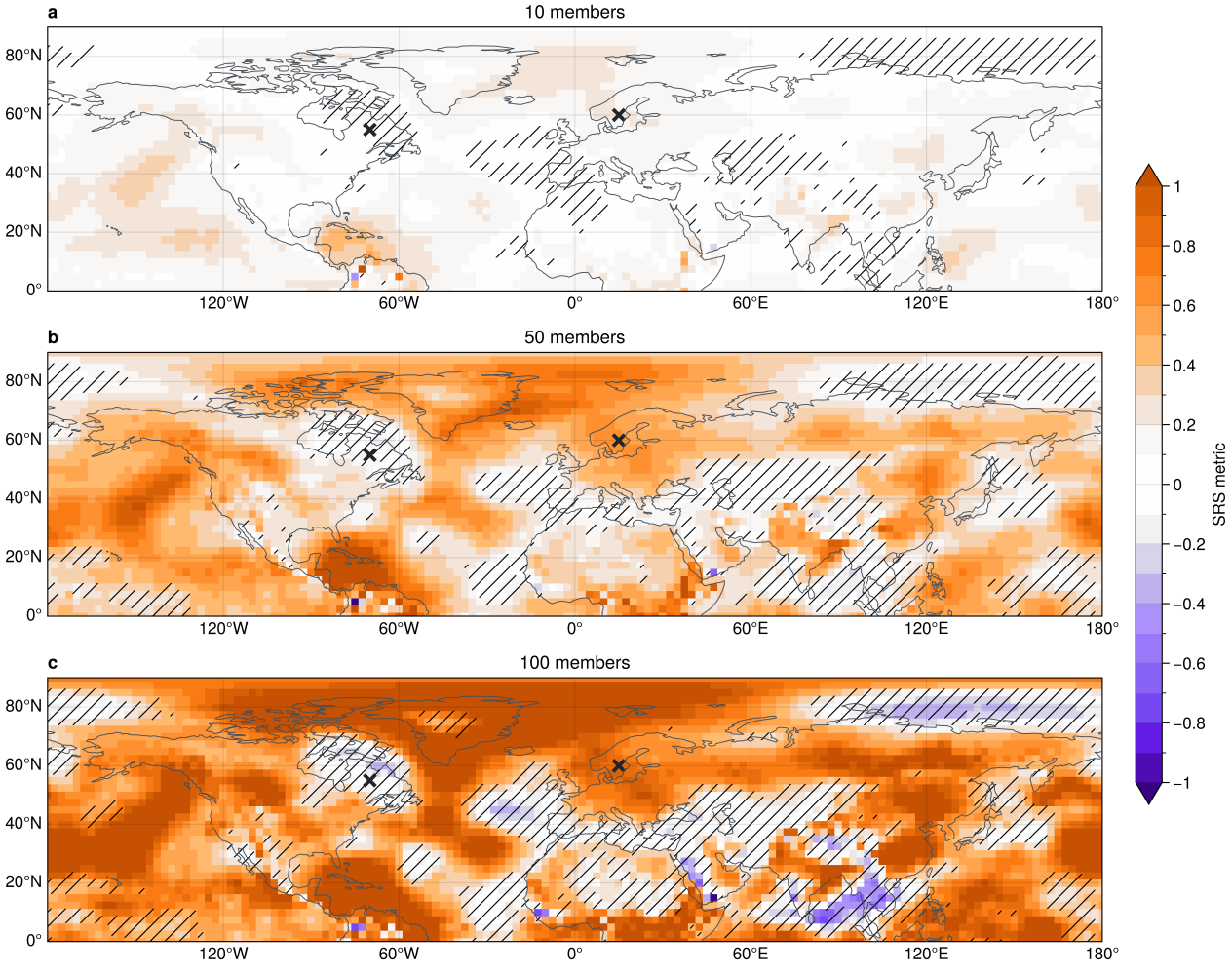

Figure 5. SRS metric computed as the slope of the z1000 spread-error curve in the northern Hemisphere. As Figure 3 but for forecast of different ensemble size (10, 50 and 100 members). Note that panel b) is identical to Fig. 3.

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

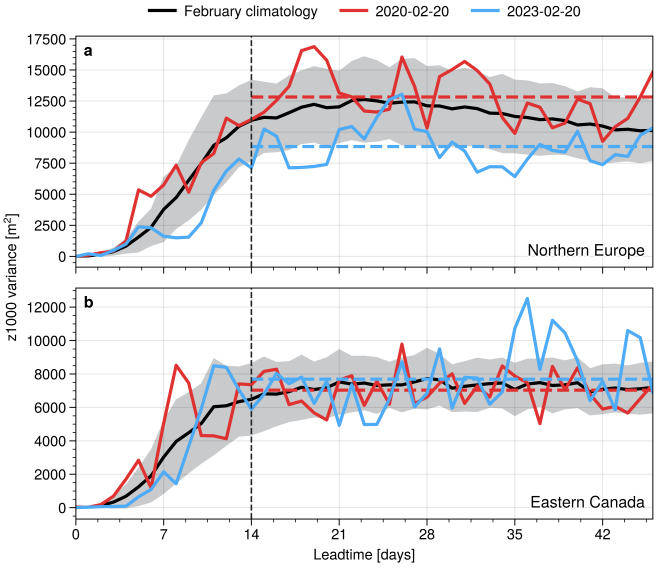

Figure 6. Evolution of leadtime-dependent ensemble variance in (a) Northern Europe at 60°N/15°E and (b) eastern Canada at 55°N/70°W. Grey line shows the climatology for February as average over all February initialisations, with shading indicating one standard deviation around the mean. Red and blue lines show variance evolution of example forecasts with 50 members initialised on February 20th of the years 2020 and 2023, respectively. Dashed horizontal lines show the average over the respective variance between days 14 and 46. Vertical dashed line indicated day 14 as visual aid.

robustly close to zero. Other effects therefore seem to play an important role here, reducing the reliability of spread fluctuations in the forecasts. In the next section, we show that a lack of variability within the physical system is a major contributor to the lack of reliability in these regions.

5.2 Intrinsic variability of the physical system

As shown within the toy model in Section 4, the intrinsic variability of spread within the underlying physical system can have a strong effect on the spread-error curve and the SRS, due to modification of the sampling error effect. To illustrate and further quantify the effect of variability in spread we contrast two example locations, one in eastern Canada (55°N/70°W) and one in northern Europe (60°N/15°E). While northern Europe shows strong improvement of SRS with increasing ensemble size and very reliable spread forecasts for 50- and 100-member ensembles, eastern Canada shows low SRS for all ensemble sizes (Fig. 5).

Figure 6 indicates the variability of ensemble spread at these two points. It can be seen that the climatological day-to-day variability in spread is generally larger in northern Europe (Fig. 5a) than eastern Canada (Fig. 5b). The difference in variability between two points in time and at a given location mostly comes from slowly varying modes of atmospheric variability that

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

affect the spread, as discussed in the following. To distinguish between slow and fast modes of variability, we will decompose changes in spread within and across forecasts into two components:

- **Inter variability**: This is due to the change in spread variability between different forecasts. It is computed as the standard deviation across forecasts of the average of ensemble variance over subseasonal leadtimes (days 14-46). It therefore describes the effect of slow modes on the evolution of the spread.
- **Intra variability**: This is due to the change in spread variability within a single forecast. It is computed as the daily standard deviation in ensemble variance over subseasonal leadtimes (days 14-46), averaged over all forecasts. It therefore describes the effect of fast modes and day-to-day variability on the spread evolution.

The concept of inter- and intra-variability is visualised in Figure 6 by two example forecasts. Both forecasts are initialised on February 20th, but in two different years: 2020 and 2023. It can be seen that in northern Europe, the two forecasts are associated with substantially different spread at subseasonal leadtimes, suggesting large inter-variability of the spread in this region. The inter-variability, i.e., the difference between subseasonally averaged variances of the two forecasts, is of the same order as the intra-variability, i.e., the day-to-day fluctuations in spread. In eastern Canada, on the other hand, the two example forecasts do essentially not differ in their subseasonal mean variance, and only show deviations from each other due to intra-variability (i.e., day-to-day variations).

Figure 7 shows the dependence of inter- and intra-variability on underlying ensemble size at the two points in eastern Canada and northern Europe. This allows us to study the two variability components more systematically. Figure 7 further shows how the two components should depend on sample size if variations were entirely due to sampling error and not due to physical drivers. It can be seen that in northern Europe the inter-variability converges clearly to a value of about 3000 m^2 for large ensembles and does not follow the theoretical line of sampling errors. This suggests a pronounced inter-variability in the physical system, which is well-sampled with ensemble sizes exceeding about 50 members. Intra variability, however, follows almost perfectly the theoretical line of sampling errors, suggesting that day-to-day variability is entirely spurious. In general, this leads to a gradual increase of the ratio of inter- over intra-variability, which exceeds one at an ensemble size of about 50.

For the point in eastern Canada, both inter- and intra-variability follow rather closely the line of sampling error theory. Even for 100 member ensembles the inter-variability has not converged yet and seems to be substantially affected by sampling errors. This leads to generally low inter-over-intra ratios. Figure 7 suggests that intra-variability is almost entirely spurious and a result of sampling error. The inter-over-intra ratio can therefore be interpreted as ratio of natural spread variability of the system compared to sampling error effects. The spatially resolved maps of inter- and intra-variability, as well as the ratio, are shown in supplementary Figure S1.

As discussed before and shown with the toy model in Section 4, large intrinsic variability can reduce the effects of sampling error on the spread-error curve and hence provide reliability of the fluctuations in ensemble spread of a forecast (i.e., increased SRS). Figure 8a shows that, indeed, regions with large inter-over-intra ratio have generally large SRS. The regions gain their spread reliability from slowly varying modes of variability that affect the forecast uncertainty. In that sense, these are also regions that show the potential to develop windows of forecast opportunity. The ensemble spread in regions with low inter-

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

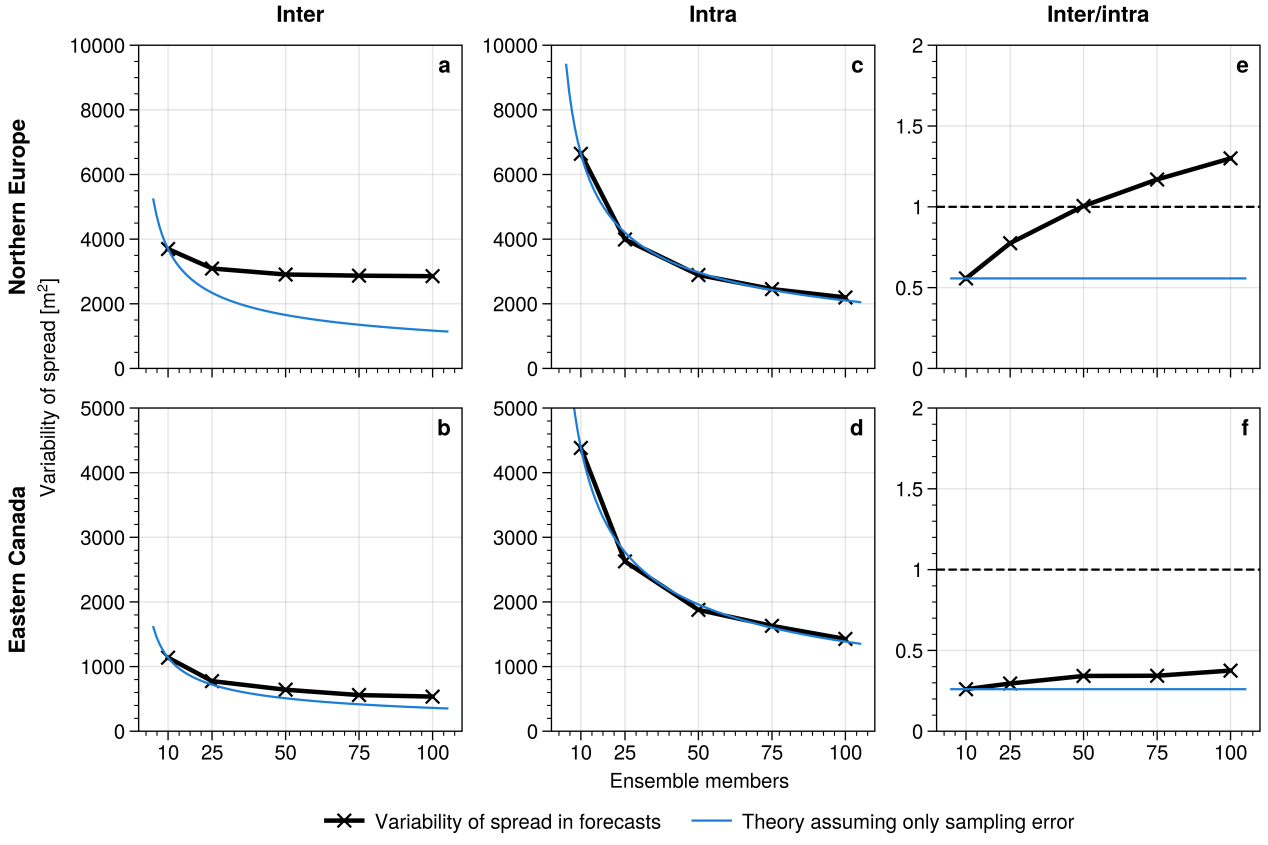

Figure 7. Average (a,b) inter and (c,d) intra variability at subseasonal leadtimes computed for forecasts with varying ensemble size. (e) and (f) show the ratio of inter-over-intra variability. Top row (a,c,e) shows a point in northern Europe at 60°N/15°E and bottom row (b,d,f) shows a point in eastern Canada at 55°N/70°W. Blue thin lines indicate theoretical dependency of variability components on ensemble size M, computed as value for 10 member ensembles divided by $\sqrt{M}$. Dashed horizontal lines in (e) and (f) indicate a ratio of 1.

over-intra ratio and corresponding low SRS (like eastern Canada) is dominated by spurious day-to-day variability but does not show robust and persistent changes in forecast uncertainty for the studied ensemble sizes. The pattern correlation between the SRS and the inter-over-intra ratio for the northern hemisphere is 0.44 based on 50-member ensembles. This correlation increases to 0.82 for the perfect model approach discussed below in Section 5.3. Further note that the regions with large inter-over-intra variability are roughly consistent with regions that show large relative variability in ensemble spread, as shown in Figure 1.

Some of the spread reliability in subseasonal z1000 spread comes from seasonal evolution, which also gives a slowly varying mode of atmospheric variability. Figure 8b shows that the slope of spread-error curves generally decreases when computed for spread and error data that has been de-seasonalised. In particular, the north Atlantic and European regions have substantially reduced spread reliability when seasonal effects are removed. Reliability at the point in northern Europe (at 60°N/15°E) reduces

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

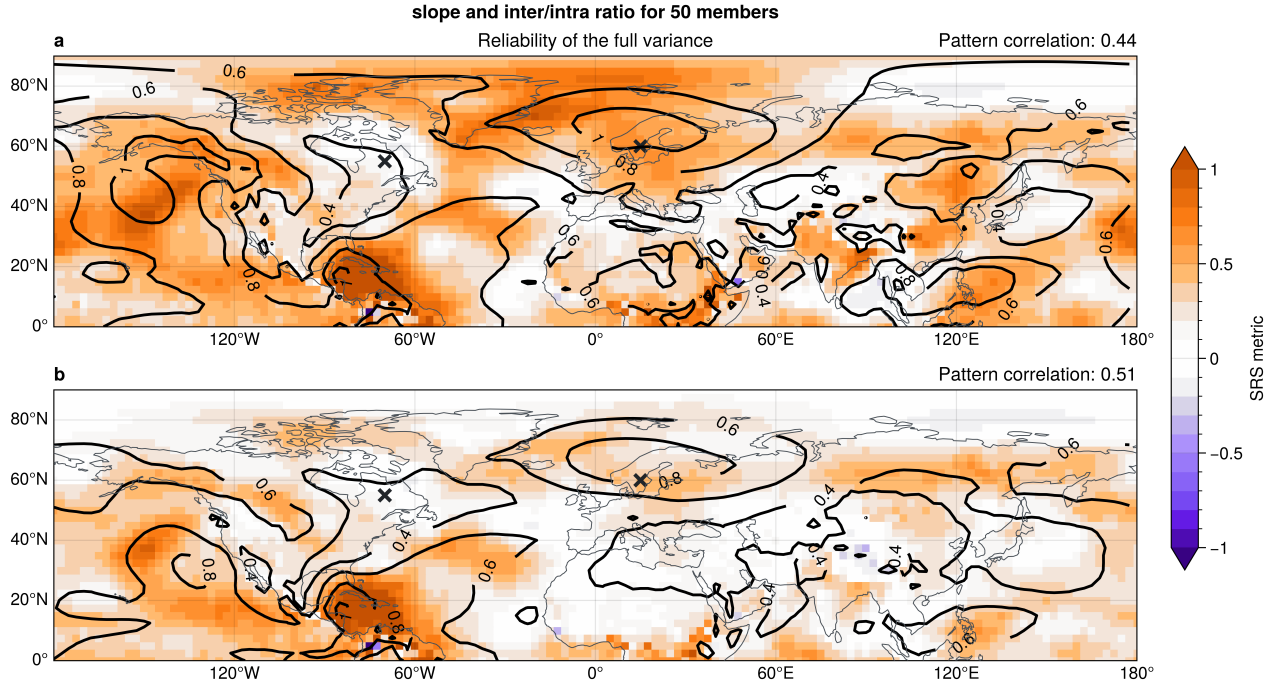

Figure 8. (shading) SRS showing the slope of z1000 spread-error curve (cf. Figure 3) and (contours) inter-over-intra variability ratio [unitless] in the northern hemisphere for 50 member ensembles. (a) computed based on the full spread and error and (b) for de-seasonalised data, where anomalies in spread and error are computed by removing a time-dependent climatology. Crosses in both panels indicate the points in northern Europe and eastern Canada analysed in other figures. Top right title in each panel shows the area-weighted pattern correlation between the slope of spread-error curves and the inter-over-intra ratio.

from 0.63 to 0.41. However, we find that inter-variability of spread is still a major source of reliability in de-seasonalised data, with pattern correlation between SRS and inter-over-intra ratio even increasing from 0.44 to 0.51 for the northern hemisphere.

5.3 Model error and under-representation of evolutions

In the previous sections we studied the effects of sampling error and natural intrinsic variability of uncertainty on the reliability of ensemble spread in subseasonal forecast models. However, forecast models may not always accurately represent all physical processes and can hence misrepresent the flow-dependence of the forecast uncertainty.

To investigate the effect of model error we performed an analysis using a perfect model approach: instead of computing the errors of the prediction as difference between the ensemble mean and re-analysis observations, we assumed the observations to be given by one of the ensemble members of the forecast. The prediction errors for an M member forecast are then given by the mean of the remaining $M-1$ ensemble members and that single selected member. The associated ensemble spread is also computed based on $M-1$ members. This approach ensures that the model spread is on average exactly equal to the prediction error, i.e., we have a perfectly reliable ensemble. However, this exact equality only holds if the approach is performed M times,

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

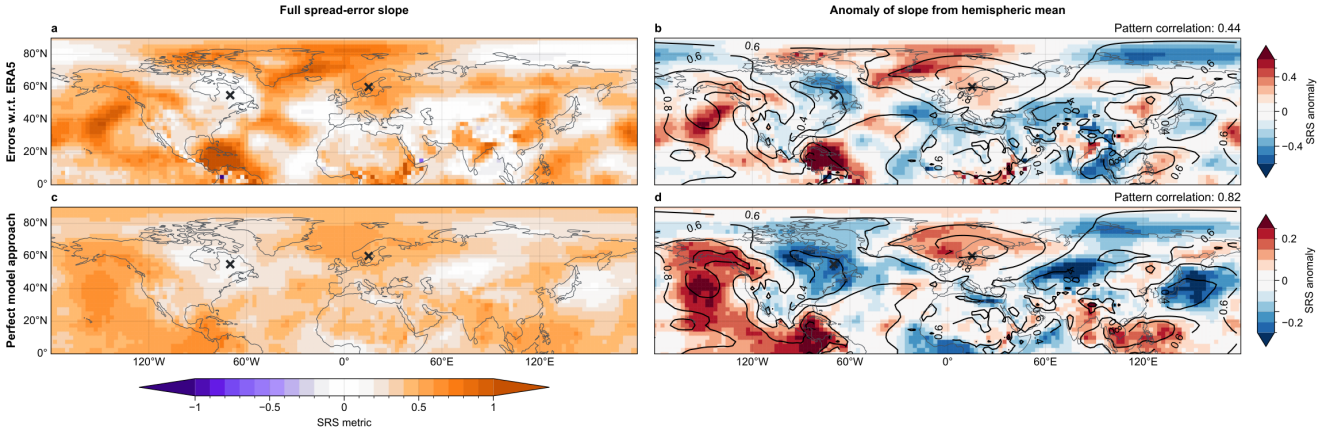

Figure 9. SRS describing the slope of z1000 spread-error curves with errors computed (a) with respect to re-analysis data and (c) based on the perfect model approach, i.e., with respect to a single ensemble member. (b) and (d) show the slope anomalies, computed as deviations from the hemispheric means, to highlight spatial structures. Contour lines in (b) and (d) show inter-over-intra ratio [unitless] with pattern correlations between slope and ratio over the northern hemisphere are indicated in the top right. All panels based on 50 member ensembles. Points in northern Europe (60°N/15°E) and eastern Canada (55°N/70°W) are indicated by crosses.

i.e., once for each of the ensemble members of the original ensemble, and then averaged over these M sets. By only computing the perfect model error and variance once for every given forecast (so based on a single "model truth"), we retain the same sampling as is given for the true errors computed based on re-analysis observations.

Figure 9 shows the SRS in the northern hemisphere for 50-member ensembles computed from re-analysis observations and using the perfect model approach, respectively. To highlight the spatial structures, Figures 9b and d show the SRS anomaly from the hemispheric mean of each respective experiment. In general, the spatial patterns in reliability of the perfect model approach matches well with the reliability computed with respect to re-analysis. This agreement indicates that the spatial patterns seen in the SRS maps mostly result from spatial inhomogeneities in the physical system (see Section 5.2). The correlation between SRS and inter-over-intra ratio for the northern hemisphere in 50-member ensembles is 0.82, further supporting the idea that spatial structures strongly result from slowly evolving modes within the underlying physical system (Figure 9d). However, magnitudes of SRS anomalies are generally smaller for the perfect model approach, suggesting that model errors in representing spread variability also play a role in modifying regional reliability.

6 Conclusions and discussion

The analysis of subseasonal winter-time forecasts, aided by an idealised toy model, shows that the reliability of ensemble spread depends on three intertwined factors: sampling error (either related to a small ensemble size or to a small number of ensemble forecast), the strength of the contribution of physically driven variability in intrinsic uncertainty, and how well the model

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

captures this variability. Regions such as northern Europe, the mid-east Pacific and the tropical west Pacific exhibit consistently high SRS values (i.e., high spread reliability), often exceeding 0.6 for 50-member ensembles. These hotspots coincide with areas influenced by slowly varying atmospheric modes that provide 'windows of forecast opportunity'. In northern Europe, for instance, the downward influence of the polar stratosphere has been linked to multi-week periods of anomalously low spread, due to a reduction in storm-induced synoptic variability. This process seems to be well-captured in forecast models and hence leads to high values of SRS in northern Europe. The mid-east Pacific signal could reflect ENSO modulation of the jet, while tropical West Pacific reliability may arise from the MJO's planetary wave response, though these connections remain speculative and warrant targeted process studies.

In contrast, eastern Canada displays almost no reliability even when 100 members are available, with SRS essentially zero. Consistently, the ensemble variance in eastern Canada is nearly constant through the subseasonal range, suggesting the atmosphere itself offers little low-frequency modulation of forecast uncertainty. Enlarging the ensemble further would therefore add computational cost without creating useful information in terms of forecast uncertainty because the intrinsic potential for windows of opportunity is vanishingly small.

The spread-error framework based on the SRS metric introduced here also serves as a diagnostic for model error. A comparison between the IFS model analysed in this study and the CNRM model (supplementary Fig. S2) shows overall similar spatial patterns of SRS despite a much smaller CNRM sample. This consistency implies that both models represent the geography of uncertainty reasonably well, although more CNRM ensemble forecasts are needed for firmer conclusions.

One practical way to exploit our findings is to define a 'corrected variance' that enforces an SRS value of one (Fig. 10). Such a correction could be constructed in various ways, with an intuitive way being the following: let $\overline{\sigma^2}$ denote the climatological mean of the ensemble spread σ^2 at a given grid point. A post-processed variance $\hat{\sigma^2}$ could be obtained based on the SRS value computed from the associated spread-error curve via $\hat{\sigma^2} = \overline{\sigma^2} + SRS^{-1}\left(\sigma^2 - \overline{\sigma^2}\right)$. This rescaling would ensure, by construction, that variance and squared error align on average. In highly reliable areas we have $SRS \approx 1$ and the correction is negligible, preserving genuine information about flow-dependent forecast skill. In unreliable areas we have $SRS \ll 1$ and the scaling damps spurious fluctuations in spread toward their climatological baseline, leading to a more trustworthy measure of uncertainty. This simple adjustment can be applied in real time and offers a transparent bridge between ensemble output and user needs for calibrated risk estimates. Note that the corrected variance $\hat{\sigma}^2$ would be close to the climatological spread in regions with low inter variability. These ideas follow closely suggestions proposed by Hopson (2014) based on ideal statistical models, where large case-to-case variability is necessary to obtain reliable and practically useful spread forecasts.

Another practical way to improve the reliability of subseasonal uncertainty estimates that emerged from this study is through time averaging. Since much of the signal originates from slowly evolving large-scale modes, averaging the daily spread over subseasonal leadtimes effectively reduces sampling noise and increases the stability of the reliability relationship. This approach mimics the effect of a larger ensemble size and can enable even 10-member ensembles to outperform the daily reliability of larger ensembles based on the SRS, as shown in supplementary Fig. S3. Alternatively, one can first average the ensemble members in time and then compute spread and error from these weekly means, which acts as a low-pass filter and emphasises slow variability. Such weekly mean datasets are widely used at subseasonal leadtimes. Supplementary Fig. S4 compares

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

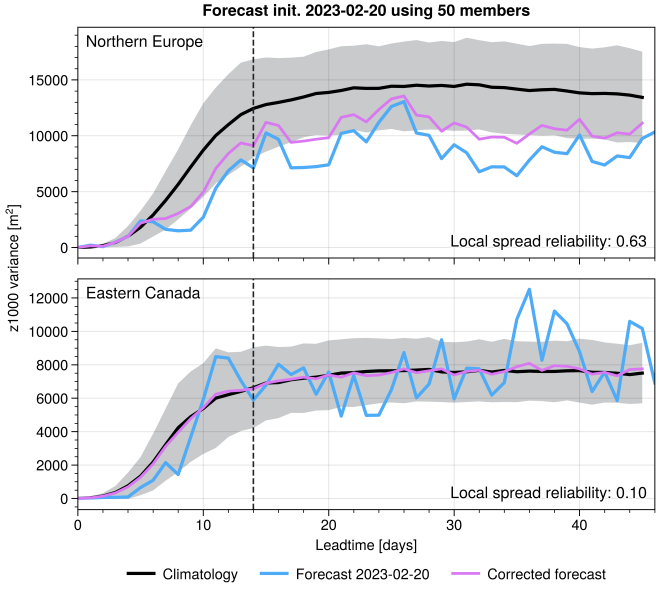

Figure 10. Evolution of z1000 ensemble variance as function leadtime at points in (a) northern Europe (60°N/15°E) and (b) eastern Canada (55°N/70°W). Black line shows climatological mean with shading indicating one standard deviation around the mean. Red line shows variance of the example forecast initialised 20th February 2023. Orange line shows the corrected ensemble variance for that example forecast by post-processing to achieve perfect reliability.

these two strategies and supports the finding of enhanced reliability through time-averaging, with subtle regional differences: for example, over the polar Atlantic, averaging the daily spread yields higher reliability (larger SRS) than computing spread from averaged fields. This difference potentially suggests that flow-dependent reliability in this region is partly linked to faster synoptic variability rather than predominantly slow modes.

While time averaging can improve reliability, other factors such as systematic model biases also influence the spread-error relationship. A systematic displacement of the ensemble mean adds a constant contribution to the squared-error term, systematically shifting each point in a spread-error scatter plot (like Fig. 2b) while leaving its slope unchanged. On the other hand, ensemble mean biases can affect the ensemble spread when internal dynamics couple the mean flow to extreme behaviour. Rupp et al. (2024), for example, discuss how an anomalous position of the Atlantic storm track in the ensemble mean flow can lower the likelihood of storms over northern Europe, thereby reducing ensemble variance in that region. For our analysis framework, however, such cases can simply be considered as model errors in terms of spread itself and should be diagnosable thorough perfect model approaches as done in Section 5.3.

Generally, the reliability of spread forecasts can vary for different variables analysed. This dependence partly arises through different flow dependences of the ensemble spread on the basic state (as, e.g., shown in Spaeth et al., 2024b). Figure 11a shows the SRS metric for the 2-metre temperature (t2m) and compares it to the t2m inter-variability. Two pronounced regions of high SRS are clearly visible around 60°N, forming band-like structures across the two major northern-hemispheric landmasses.

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

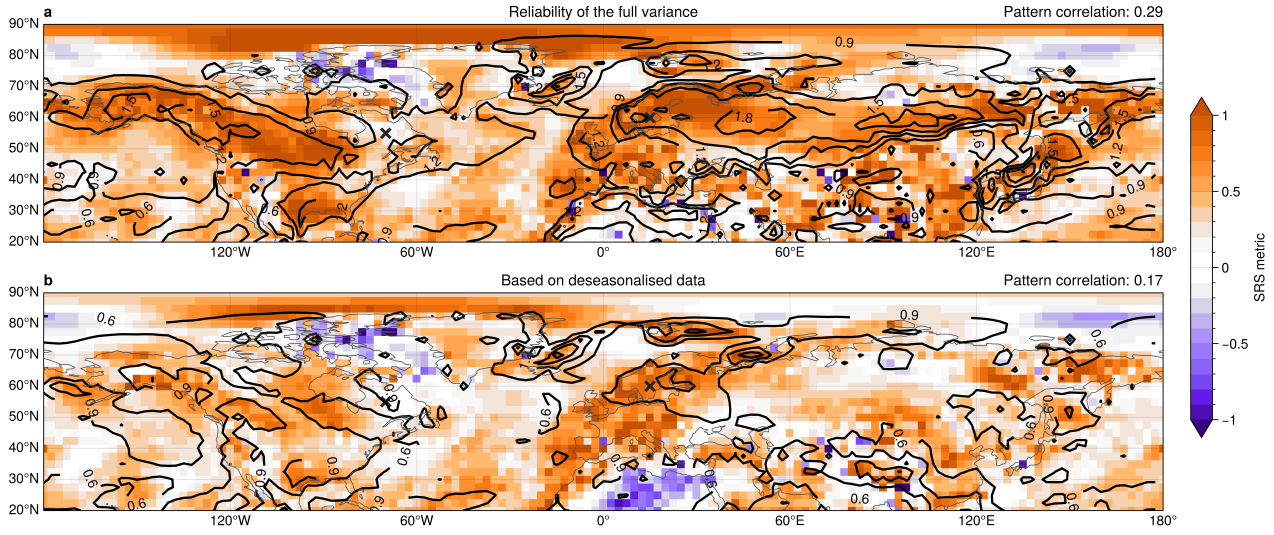

Figure 11. As Fig. 8 but for t2m rather than z1000.

These regions also show high values of inter-over-intra variability, further suggesting that subseasonal reliability is mostly driven by slowly evolving modes, although the overall pattern correlation for the northern hemisphere is relatively small. Further, we find that the high SRS for the t2m field in these regions mostly arises from the seasonal winter-to-spring transition in surface temperatures, which is most pronounced in the mid-latitudes and over land. Correspondingly, the SRS strongly decreases in those regions when computing the reliability based on deseasonalised spread data (Figure 11b).

The results presented here might also be relevant with regard to the so-called 'signal-to-noise paradox' (SNP), described by Scaife and Smith (2018). The paradox refers to an apparent mismatch in climate and seasonal forecasting systems, in which forecasts correlate better with observed variability than with their own ensemble members. According to Scaife and Smith (2018), such a situation arises if the unpredictable component (noise) of the observed atmosphere is systematically smaller than the ensemble spread suggests, leading to forecasts being paradoxically under-confident. Related work by Strommen et al. (2023) further suggests that the SNP is equivalent to reliability diagrams exhibiting slopes greater than unity, although note that their reliability metric differs from our spread-error slopes. Despite methodological differences, both the SNP and our spread-error analysis highlight the fundamental importance of accurately capturing atmospheric variability in ensemble spread.

While our framework does not directly quantify the absolute level of ensemble spread (which is central to the SNP), it does address whether fluctuations in spread (e.g., departures from climatology) reliably represent variations in atmospheric uncertainty. Regions identified in our study as having low inter-over-intra variability ratios also exhibit correspondingly poor spread reliability and low SRS. Such conditions might indirectly reflect scenarios favourable for the paradox, potentially arising from systematic misrepresentations of ensemble spread variability. Indeed, Karpechko et al. (2025) demonstrate that teleconnections influencing the subseasonal skill are also associated with changes in the ensemble spread, further supporting a potential linkage between slowly varying atmospheric modes, ensemble spread representation, and the SNP. A more direct

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

465 analysis of how precisely variability in spread connects to the paradox remains an open question and future research explicitly bridging these concepts could help clarify their relationship.

In summary, the ability of an ensemble to convey reliable uncertainty forecasts depends on two questions: does the physical system provide a window of opportunity, and is the model accurate enough to detect it? Our spread-error framework shows that, over large areas of the Northern Hemisphere, those windows are opened by slowly varying teleconnections and can already be
470 well-resolved with 50 to 100 members, whereas regions lacking a strong influence of such slow modes remain unreliable even when the ensemble size is large.

Acknowledgements. The authors thank the Transregional Collaborative Research Center SFB/TRR 165 "Waves to Weather" funded by the German Research Foundation (DFG) for support. We further thank Hella Garny for some inspirational discussions about predictability.

Author contributions. PR and JS conceptualised the idea together. PR performed the analyses of subseasonal forecasts and wrote most of
475 the manuscript. JS performed the toy model simulations and wrote the corresponding section. TB assisted with the interpretation of results and helped to improve the manuscript.

Competing interests. The authors declare that they have no conflict of interest. Thomas Birner is a member of the editorial board of Weather and Climate Dynamics.

Data availability. Detailed information on the datasets used can be found in Section 2.

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

References

Baggett, C. F., Barnes, E. A., Maloney, E. D., and Mundhenk, B. D.: Advancing atmospheric river forecasts into subseasonal-to-seasonal time scales, Geophysical Research Letters, 44, 7528–7536, 2017.

Baldwin, M. P., Stephenson, D. B., Thompson, D. W., Dunkerton, T. J., Charlton, A. J., and O'Neill, A.: Stratospheric memory and skill of extended-range weather forecasts, Science, 301, 636–640, 2003.

Bröcker, J. and Smith, L. A.: Increasing the reliability of reliability diagrams, Weather and forecasting, 22, 651–661, 2007.

Fortin, V., Abaza, M., Anctil, F., and Turcotte, R.: Why should ensemble spread match the RMSE of the ensemble mean?, Journal of Hydrometeorology, 15, 1708–1713, 2014.

Giggins, B. and Gottwald, G. A.: Stochastically perturbed bred vectors in multi-scale systems, Quarterly Journal of the Royal Meteorological Society, 145, 642–658, 2019.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly journal of the royal meteorological society, 146, 1999–2049, 2020.

Hopson, T.: Assessing the ensemble spread–error relationship, Monthly Weather Review, 142, 1125–1142, 2014.

Johnson, N. C., Collins, D. C., Feldstein, S. B., L'Heureux, M. L., and Riddle, E. E.: Skillful wintertime North American temperature forecasts out to 4 weeks based on the state of ENSO and the MJO, Weather and Forecasting, 29, 23–38, 2014.

Karpechko, A. Y., Butler, A. H., and Vitart, F.: Signal, noise and skill in sub-seasonal forecasts: the role of teleconnections, EGUsphere, 2025, 1–30, 2025.

Lakatos, M., Lerch, S., Hemri, S., and Baran, S.: Comparison of multivariate post-processing methods using global ECMWF ensemble forecasts, Quarterly Journal of the Royal Meteorological Society, 149, 856–877, 2023.

Leutbecher, M. and Palmer, T. N.: Ensemble forecasting, Journal of computational physics, 227, 3515–3539, 2008.

Leutbecher, M., Lang, S., Lock, S.-J., Roberts, C. D., and Tsiringakis, A.: Improving the physical consistency of ensemble forecasts by using SPP in the IFS.

Mariotti, A., Baggett, C., Barnes, E. A., Becker, E., Butler, A., Collins, D. C., Dirmeyer, P. A., Ferranti, L., Johnson, N. C., Jones, J., et al.: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond, Bulletin of the American Meteorological Society, 101, E608–E625, 2020.

Robertson, A. W., Vigaud, N., Yuan, J., and Tippett, M. K.: Toward identifying subseasonal forecasts of opportunity using North American weather regimes, Monthly Weather Review, 148, 1861–1875, 2020.

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.

Scaife, A. A. and Smith, D.: A signal-to-noise paradox in climate science, npj Climate and Atmospheric Science, 1, 28, 2018.

Scherrer, S. C., Appenzeller, C., Eckert, P., and Cattani, D.: Analysis of the spread–skill relations using the ECMWF ensemble prediction system over Europe, Weather and Forecasting, 19, 552–565, 2004.

Selz, T.: Estimating the intrinsic limit of predictability using a stochastic convection scheme, Journal of the Atmospheric Sciences, 76, 757–765, 2019.

Selz, T., Riemer, M., and Craig, G. C.: The transition from practical to intrinsic predictability of midlatitude weather, Journal of the Atmospheric Sciences, 79, 2013–2030, 2022.

https://doi.org/10.5194/egusphere-2025-4925
Preprint. Discussion started: 23 October 2025

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.

Spaeth, J., Rupp, P., Osman, M., Grams, C., and Birner, T.: Flow-dependence of ensemble spread of subseasonal forecasts explored via North Atlantic-European weather regimes, Geophysical Research Letters, 51, e2024GL109 733, 2024b.

Strommen, K., MacRae, M., and Christensen, H.: On the Relationship Between Reliability Diagrams and the "Signal-To-Noise Paradox", Geophysical Research Letters, 50, e2023GL103 710, 2023.

Tempest, K. I., Craig, G. C., and Brehmer, J. R.: Convergence of forecast distributions in a 100,000-member idealised convective-scale ensemble, Quarterly Journal of the Royal Meteorological Society, 149, 677–702, 2023.

Vitart, F. and Robertson, A. W.: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events, npj climate and atmospheric science, 1, 3, 2018.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., et al.: The subseasonal to seasonal (S2S) prediction project database, Bulletin of the American Meteorological Society, 98, 163–173, 2017.

Weisheimer, A. and Palmer, T.: On the reliability of seasonal climate forecasts, Journal of the Royal Society Interface, 11, 20131 162, 2014.

Weisheimer, A., Decremer, D., MacLeod, D., O'Reilly, C., Stockdale, T. N., Johnson, S., and Palmer, T. N.: How confident are predictability estimates of the winter North Atlantic Oscillation?, Quarterly Journal of the Royal Meteorological Society, 145, 140–159, 2019.