A statistical perspective on the signal–to–noise paradox

Article

Published Version

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

To link to this article DOI: http://dx.doi.org/10.1002/qj.4440

Publisher: Royal Meteorological Society

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

RESEARCH ARTICLE

A statistical perspective on the signal-to-noise paradox

Jochen Bröcker[1,2] | **Andrew J. Charlton–Perez**[1,2] | **Antje Weisheimer**[3,4]

[1]Department of Mathematics and Statistics, Centre for Mathematics of Planet Earth, University of Reading, Reading, UK

[2]Department of Meteorology, University of Reading, Reading, UK

[3]European Centre for Medium Range Weather Forecasts, Reading, UK

[4]Department of Physics, National Centre for Atmospheric Science (NCAS), University of Oxford, Oxford, UK

Correspondence
J. Bröcker, Department of Mathematics and Statistics, University of Reading, Whiteknights, Reading, Berkshire RG6 6AX, UK.
Email: j.broecker@reading.ac.uk

Abstract

An anomalous signal-to-noise ratio (also called the signal-to-noise paradox) present in climate models has been widely reported, affecting predictions and projections from seasonal to centennial timescales and encompassing prediction skill from internal processes and external climate forcing. An anomalous signal-to-noise ratio describes a situation where the mean of a forecast ensemble correlates better with the corresponding verification than with its individual ensemble members. This situation has severe implications for climate science, meaning that large ensembles might be required to extract prediction signals. Although a number of possible physical mechanisms for this paradox have been proposed, none has been universally accepted. From a statistical point of view, an anomalous signal-to-noise ratio indicates that forecast ensemble members are not statistically interchangeable with the verification, and an apparent paradox arises only if such an interchangeability is assumed. It will be demonstrated in this study that an anomalous signal-to-noise ratio is a consequence of the relative magnitudes of the variance of the observations, the ensemble mean, and the error of the ensemble mean. By analysing the geometric triangle formed by these three quantities, and given that for typical seasonal forecasting systems both the correlation and the forecast signal are relatively small, it is concluded that an anomalous signal-to-noise ratio should, in fact, be expected in such circumstances.

KEYWORDS

climate model forcing, ensemble forecasts, signal-to-noise ratio

1 | INTRODUCTION

Modern dynamical medium-range and seasonal-scale weather forecasting models are self-assessing in that they not only report a single ("best guess" or "most probable") scenario for the future evolution of the weather, but in addition provide statements regarding the predicted accuracy of their own forecasts (or, equivalently, the associated uncertainty). In many operational forecasting systems this is accomplished through ensemble forecasts (see for instance Leutbecher and Palmer, 2008; Weigel, 2011); a complete instance of the forecast comprises not one but several (typically between 10 and 100) scenarios for the future evolution of the weather. These scenarios are gen-

erated by perturbations to the initial state, model parameters, or added stochastic noise. Typically, each scenario (or ensemble member) is deemed equally likely given the information available about the state of the atmosphere, which needs to be determined from inherently uncertain meteorological observations. More generally, we may think of modern forecasting systems as providing a very rough approximation of the distribution of the future, given the information available and uncertainties present at the time the forecast is issued.

It should not come as a surprise that many such forecasting systems turn out to be overconfident; the uncertainties predicted by the forecasting system are smaller than the actual uncertainties (Weisheimer *et al.,* 2011; Weisheimer and Palmer, 2014). The overconfidence of ensemble forecasts is a major motivation for the development of stochastic physics approaches to represent model uncertainties in dynamical forecast models explicitly (Palmer, 2019). In the case of ensemble forecasts, for instance, the squared distance between the ensemble mean and individual ensemble members should, when averaged over all ensemble members and over time, be a good estimate of the distance between the ensemble mean and the actual verification (again averaged over time, Palmer *et al.,* 2006). The former distance is often significantly smaller than the latter, leading to overconfident forecasts.

Recently, several studies have reported instances of forecasting systems that appear to be *underconfident* in a certain sense (Eade *et al.,* 2014; Scaife *et al.,* 2014; Stockdale *et al.,* 2015; Baker *et al.,* 2018; Scaife and Smith, 2018; Charlton-Perez *et al.,* 2019; Weisheimer *et al.,* 2019) Here, we do *not* mean that the errors predicted by the forecasting system are larger than the actual errors obtained when comparing those forecasts with verifications. Rather, the ensemble mean forecast of these systems correlates better with the verification than would be expected by comparing the ensemble mean forecast with another realisation from the forecasting system. Therefore, those forecasts seem to contain more information when correlated with the real world than they *pretend* to contain (by their self-assessing nature) when correlated against the model world. This phenomenon, first reported in Scaife *et al.* (2014) (and mentioned in Kumar, 2009, as a possibility), has been termed the *signal-to-noise paradox.* The review of Scaife and Smith (2018) shows the effect is present for forecasts on a range of timescales, including not just initialised seasonal and decadal predictions but also the response of climate models to external forcing such as large volcanic eruptions. If climate models exhibit a signal-to-noise paradox, this has broad and far-reaching consequences for our ability to use them for making useful climate predictions and projections. In particular, it

implies that running large ensembles will be necessary to isolate robust climate signals, constraining the ways in which finite computational resources can be used.

In the present work, it is argued that, depending on the precise circumstances, the signal-to-noise paradox should not actually appear paradoxical. The ensemble dispersion is directly related to the correlation of the ensemble mean with individual ensemble members, and the behaviour described above appears paradoxical only if the forecast error is assumed to be similarly related to the correlation of the ensemble mean with the verification, which is not the case. This remains true even if the ensemble dispersion and mean-squared forecast error agree, as they typically do, at least approximately. For instance, if the forecast or the verification is multiplied by a factor, or if a constant is added to either, the correlation between them does not change. The forecast error, however, does change and thus cannot be determined by the correlation alone. Examples of this effect are widespread in the literature: for example, figure 1 of Scaife *et al.* (2014) and figure 1 of Stockdale *et al.* (2015). In both cases, for ease of comparison with the observed state, a series of ensemble forecasts is scaled by a constant value. This does not change the correlation between the ensemble mean and the observations, but does change the properties of the forecast error (see also Smith *et al.,* 2020).

Fundamentally, the signal-to-noise paradox in an ensemble forecasting system indicates that the ensemble members and verification are not statistically equivalent, which can be regarded as a lack of reliability or calibration (Bröcker, 2021). Whether this is caused by a single underlying physical phenomenon is currently not known, yet it is conceivable that, even in a single forecasting system exhibiting a signal-to-noise paradox, more than one effect might contribute. Scaife and Smith (2018) and Smith *et al.* (2020) discuss the weak response to external forcings (such as volcano aerosols, solar variability, Arctic sea-ice loss, and the dynamics of the ozone hole) as a possible reason. In Zhang *et al.* (2021), poor representation of ocean mesoscale processes is identified as a partial explanation of the signal-to-noise paradox, especially over eddy-rich regions of the oceans. Markov-type models are used in Strommen and Palmer (2019) and Zhang and Kirtman (2019) to link the signal-to-noise paradox to underestimated persistence and regime behaviour, while Hardiman *et al.* (2022) identify missing eddy feedback as a physical mechanism. Finally, evidence for a signal-to-noise paradox in seasonal forecasts involving the stratosphere is mixed, with ongoing discussions regarding whether and how the stratosphere adds predictive skill (Seviour *et al.,* 2014; Byrne *et al.,* 2019; Charlton-Perez *et al.,* 2019; O'Reilly *et al.,* 2019).

Although identifying the underlying reasons is extremely important given the implications, they are not the focus of the present contribution. Rather, we will present an analysis of the signal-to-noise paradox from a statistical perspective. Our findings are in line with earlier studies (Eade *et al.,* 2014; Siegert *et al.,* 2016; Scaife and Smith, 2018; Christiansen *et al.,* 2022), although a different and more quantitative angle allows us to derive several complementary results. Furthermore, our analysis does not rely on assuming a specific statistical model as in Siegert *et al.* (2016), Charlton-Perez *et al.* (2019), and Christiansen *et al.* (2022).

The structure of this article is as follows. In Section 2 the correlation coefficient between a generic forecast and a corresponding verification will be formally introduced and discussed, along with its connection to the *ratio of predictable components* (RPC: Eade *et al.,* 2014). The role of the correlation coefficient as a measure of information content (broadly speaking) will be contrasted with the mean-squared error as a measure of actual forecast performance, and we will define the concepts of a normal and anomalous signal-to-noise ratio. The approach allows us to consider these concepts for generic (possibly deterministic) forecasts. The discussion in Section 3 focusses on ensemble forecasts. More specifically, the ensemble mean takes the role of the forecast. In this setup, we analyse the situation of normal and anomalous signal-to-noise ratios, and in particular provide reasons as to why an anomalous signal-to-noise ratio might seem paradoxical. Section 4 will discuss aspects of the performance of seasonal ensemble forecasting systems of the European Centre for Medium-Range Weather Forecasts (ECMWF), namely the operational SEAS5 system as well as the lower resolution Coupled Seasonal Forecasts of the 20th Century (CSF–20C) system, along with uncoupled (i.e., atmosphere only, ASF–20C) versions of these systems (see Weisheimer *et al.,* 2020). The observations of previous studies, in particular with regards to the signal-to-noise paradox, will be revisited in the light of the discussion in previous sections. Conclusions and avenues for future work will be presented in Section 5.

2 | THE RATIO OF PREDICTABLE COMPONENTS AND THE SIGNAL-TO-NOISE RATIO

We fix some notation first, which we will use throughout the article. For a general random variable Z, we write $\mathbb{E}(Z)$ and $\mathbb{V}(Z) := \mathbb{E}\left[(Z - \mathbb{E}(Z))^2\right]$ for the expected value and the variance of Z, respectively. Further,

$$\text{Cov}(Z_1, Z_2) := \mathbb{E}\left[(Z_1 - \mathbb{E}(Z_1))(Z_2 - \mathbb{E}(Z_2))\right]$$

denotes the covariance between two random variables Z_1 and Z_2. It will be convenient to assume that all random variables have mean zero, or $\mathbb{E}(Z) = 0$. This convention implies that the forecasts and verifications have been rendered mean-free by subtracting the empirical mean from both; note, however, that the empirical means of forecasts and verifications are not necessarily the same, so this step corresponds to a simple postprocessing or "debiasing" of the forecasts, which is common practice in seasonal forecasting. With this assumption, we have that $\mathbb{V}(Z) = \mathbb{E}(Z^2)$. Therefore the root of the variance $\sqrt{\mathbb{V}(Z)}$ of a random variable can be interpreted as the "magnitude" or "amplitude" of that random variable.

For the moment, we consider a single verification (or observation) Y as a random variable with values that are real numbers. The corresponding forecast X is likewise a real-valued random variable (ensemble forecasts will be considered in Section 3).

We recall the Pearson correlation coefficient

$$\rho := \frac{\text{Cov}(X, Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}} = \frac{\mathbb{E}(XY)}{\sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}},$$

where the second version applies to mean-free random variables (recall that this is our convention).

The correlation coefficient can be interpreted in the following way. Consider a triangle with sides of length $\sqrt{\mathbb{V}(X)}$, $\sqrt{\mathbb{V}(Y)}$, and $\sqrt{\mathbb{V}(X - Y)}$, with α being the angle between the first two sides (Figure 1). Now, on the one hand, using a standard trigonometric identity,[1] we find

$$\mathbb{V}(Y - X) = \mathbb{V}(X) + \mathbb{V}(Y) - 2\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}\cos(\alpha). \quad (1)$$

On the other hand, by expanding the left-hand side we find

$$\mathbb{V}(Y - X) = \mathbb{V}(X) + \mathbb{V}(Y) - 2\mathbb{E}(XY).$$

Comparing the two previous relations, we obtain

$$\cos(\alpha) = \rho.$$

The concepts of normal and anomalous signal-to-noise ratios (to be defined below) are typically introduced in the context of ensemble forecasts (Eade *et al.,* 2014; Siegert

---

[1] To see this, write $\sqrt{\mathbb{V}(X)} = x_1 + x_2$, where x_1 and x_2 split the line $\sqrt{\mathbb{V}(X)}$ at the intersection point with height h. Since h divides the triangle into two right-angled triangles, we have, by Pythagoras' theorem, $h^2 = \mathbb{V}(Y - X) - x_2^2 = \mathbb{V}(Y) - x_1^2$, so $\mathbb{V}(Y - X) = \mathbb{V}(Y) + x_2^2 - x_1^2$, and since $x_2 = \sqrt{\mathbb{V}(X)} - x_1$ we have $\mathbb{V}(Y - X) = \mathbb{V}(Y) + \mathbb{V}(X) - 2\sqrt{\mathbb{V}(X)}x_1$. From this relation we can eliminate x_1 with the identity $\cos(\alpha) = x_1/\sqrt{V(Y)}$, and Equation (1) emerges.
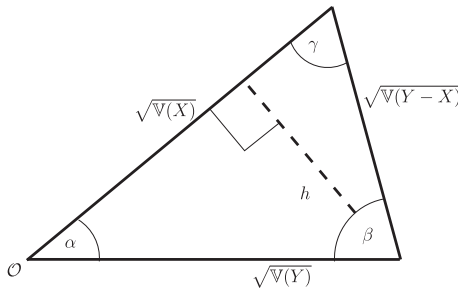
FIGURE 1 A triangle illustrating the interpretation of the correlation coefficient as $\cos(\alpha)$ (see also Rodgers and Nicewander, 1988).
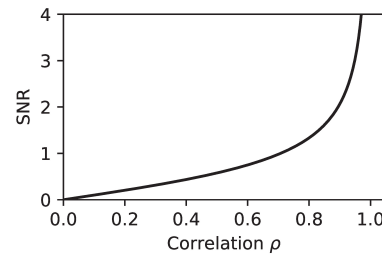

FIGURE 2 The relationship between the correlation coefficient (abscissa) and the signal-to-noise ratio (ordinate). See also figure 2 in Kumar (2009).

et al., 2016; Scaife and Smith, 2018). They can be defined for deterministic forecasts as well, however, by comparing the actual correlation ρ with the correlation obtained if the forecast X has a certain optimality property, namely that the triangle formed by $\sqrt{\mathbb{V}(X)}$, $\sqrt{\mathbb{V}(Y)}$, and $\sqrt{\mathbb{V}(Y-X)}$ in Figure 1 is right-angled (with $\sqrt{\mathbb{V}(Y)}$ forming the hypotenuse and γ a right angle). If this is the case, we will refer to the forecasts as being *linearly calibrated*, for reasons that will become clear soon. The forecast X is linearly calibrated if and only if

$$\mathbb{V}(Y) = \mathbb{V}(X) + \mathbb{V}(Y-X), \tag{2}$$

which is a consequence of Pythagoras' theorem. If the forecast is linearly calibrated, the correlation coefficient ρ is given by the quantity

$$\rho_f := \sqrt{\frac{\mathbb{V}(X)}{\mathbb{V}(X) + \mathbb{V}(Y-X)}}, \tag{3}$$

which we call the *model correlation coefficient* (a term we will use no matter whether the forecast is linearly calibrated or not). To see this, use Equation (2) to eliminate $\mathbb{V}(Y)$ in Equation (1) and solve for $\cos(\alpha) = \rho$. Note that rescaling of X will not change either Y or the angle α (or equivalently ρ), but it will change the error variance $\mathbb{V}(X-Y)$; it is easy to see that the error is minimal if the forecast X is linearly calibrated. Said differently, if the forecast X is linearly calibrated, then any rescaling of X (while keeping ρ and Y fixed) would make the error $\mathbb{V}(X-Y)$ larger. As a consequence, the mean-square error of a linearly calibrated forecast cannot be improved by linear rescaling of the forecasts.

The ratio RPC $:= \rho/\rho_f$ will be referred to as the *ratio of predictable components*. This quantity was introduced in Eade *et al.* (2014), albeit with ρ_f defined in the context of ensemble forecasts. (More specifically, ρ_f will be replaced with the correlation between the ensemble mean and individual ensemble members.) This will be

discussed in Section 3 in more detail, where it should also become clearer why we call ρ_f the "model correlation coefficient". If the forecasts are linearly calibrated, then $\rho_f = \rho$ as discussed and therefore RPC = 1. Following Scaife and Smith (2018), we introduce the following definition.

> **Definition 1.** We will say that the verification–forecast pair (Y, X) exhibits an *anomalous signal-to-noise ratio* if RPC > 1. Otherwise, we will say that the signal-to-noise ratio is *normal*.

Given the previous discussion, it would probably be more appropriate to speak of the normal and anomalous ratio of predictable components. The term "signal-to-noise ratio" was chosen, however, to link clearly to the prior literature on this topic. The quantity SNR = $\cot(\alpha)$ can be interpreted as the signal-to-noise ratio, but provides the same information as the correlation coefficient $\rho = \cos(\alpha)$ (see Figure 2, which is figure 2 in Kumar (2009) with axes interchanged), and we will continue working with the latter. We will also use the terms "normal RPC" (respectively "anomalous RPC") synonymously with "normal SNR" (respectively "anomalous SNR").

We stress that our approach permits us to define the RPC as well as normal and anomalous signal-to-noise ratios without having to resort to more explicit statistical models, such as, for instance, in Siegert *et al.* (2016), Charlton-Perez *et al.* (2019), and Christiansen *et al.* (2022). Nonetheless, it needs to be kept in mind that the RPC is expressed in terms of variances and covariances and therefore has to be estimated; any estimated RPC will clearly carry uncertainties. These uncertainties will bear on whether a forecasting system is classified as exhibiting a normal or an anomalous signal-to-noise ratio. In other words, whether a forecasting system exhibits a normal or an anomalous signal-to-noise ratio requires a statistical test. Developing such tests will be subject to future research.

3 | ANOMALOUS SIGNAL-TO-NOISE RATIOS AND ENSEMBLE FORECASTS

In this section, we hope to clarify a point regarding the signal-to-noise ratio that has generated considerable discussion in the literature. Most of this discussion centres around the performance of seasonal or decadal ensemble prediction systems. An ensemble can be considered as a set $(X_1, \dots, X_K)$ of forecasts for the same verification Y. Each ensemble member X_k represents a possible value for Y; in an ideal ensemble, the members $(X_1, \dots, X_K)$ are "statistically indistinguishable from Y" given the currently available information such as the current state of the atmosphere. This property is described adequately mathematically (see Bröcker and Kantz, 2011) by assuming the joint distribution of ensemble members and the verification to be exchangeable, that is, the distribution of $(Y, X_1, \dots, X_K)$ is symmetric.

The ensemble mean

$$X = \frac{1}{K} \sum_{k=1}^{K} X_k$$

may then be used as a forecast, and in particular the RPC for this forecast can be considered. In Weisheimer *et al.* (2019), for instance, the performance of several different ECMWF seasonal ensemble prediction systems (System 4, SEAS5, and ASF-20C) is compared over the common hindcast period 1981–2009. Over this period, the ensemble mean shows an anomalous signal-to-noise ratio for certain regions, in agreement with other studies.

We stress that in Eade *et al.* (2014), Weisheimer *et al.* (2019), and also elsewhere the model correlation coefficient ρ_f is defined with the ensemble dispersion in the denominator rather than the mean-square error $\mathbb{V}(X - Y)$. More precisely, instead of ρ_f, the quantity

$$\rho_\sigma = \sqrt{\frac{\mathbb{V}(X)}{\mathbb{V}(X) + \sigma^2}}$$

is used in lieu of the model correlation coefficient, with the ensemble dispersion defined as $\sigma^2 = \mathbb{V}(X_k - X)$. The ensemble dispersion is the same for all k, as the ensemble members are supposed to be exchangeable. It can be estimated by averaging $(X_k - X)^2$ over all ensemble members and all forecast instances. Using ρ_σ instead of ρ_f in the definition of the ratio of predictable components, we define $\text{RPC}_\sigma := \rho/\rho_\sigma$. We will refer to ρ_σ as the *ensemble correlation coefficient* in order to distinguish it from the model correlation coefficient ρ_f. The

ensemble correlation coefficient ρ_σ provides the correlation between the ensemble mean and the individual ensemble members.

In general, RPC and RPC_σ are different and, in particular, one of these might be anomalous while the other is not. In the case of the atmospheric seasonal hindcasts discussed in Weisheimer *et al.* (2019), however, over the 110-year period from 1900–2009 the dispersion in fact matches the mean-square error $\mathbb{V}(X - Y)$ of the ensemble mean approximately (see figure 7h in that article). Therefore, for the experiments in Weisheimer *et al.* (2019), the values of ρ_f and RPC, respectively, will be very similar to those of ρ_σ and RPC_σ, respectively.

As was already mentioned in the Introduction, typical seasonal forecast ensembles are underdispersive, in the sense that $\sigma^2 < \mathbb{V}(Y - X)$ (this includes the systems considered in Section 4). All other things being equal, this implies that $\rho_f < \rho_\sigma$ and hence $\text{RPC} > \text{RPC}_\sigma$. Hence an anomalous signal-to-noise ratio according to RPC_σ will then also be anomalous with respect to RPC. In our numerical experiments, we will consider both quantities, which however provide the same conclusions. It is also worth stressing that the finite ensemble size leads to an overestimate of $\mathbb{V}(X)$ and therefore an underestimate of ρ due to residual noise, with the consequence that finite ensembles tend to underestimate both RPC and RPC_σ (Scaife and Smith, 2018).

3.1 | Why an anomalous SNR might appear paradoxical

Let us assume for the moment that the ensemble dispersion σ^2 is equal to the mean-square error $\mathbb{V}(Y - X)$ of the ensemble mean (a property that is sometimes called *marginal calibration*). In that situation, an anomalous signal-to-noise ratio may lead to the apparently paradoxical conclusion that the ensemble mean is better in predicting the verification than in predicting individual ensemble members. Here, "better in predicting" means "being more strongly correlated with". This conclusion is correct, as we will see now, but, as we will argue in the next subsection, appears paradoxical only if unwarranted assumptions are made regarding the statistical equivalence (or interchangeability) of ensemble members and verifications. An anomalous signal-to-noise ratio means that $\rho > \rho_\sigma$, and we will now demonstrate that ρ_σ is indeed the correlation between the ensemble mean X and the individual ensemble members X_k (justifying us calling this quantity the "ensemble correlation coefficient"). To see this, we consider the quantity $\mathbb{E}(X_k X)$, and note that it does not depend on k, due to our assumption that the ensemble members are exchangeable.

Therefore it will not change if we average over k. This gives

$$\mathbb{E}(X_k X) = \frac{1}{K} \sum_{l=1}^{K} \mathbb{E}(X_l X) = \mathbb{E}\left(\frac{1}{K} \sum_{l=1}^{K} X_l X\right)$$
$$= \mathbb{E}(X^2) = \mathbb{V}(X). \qquad (4)$$

This, however, implies that in the following calculation the $\mathbb{E}(\dots)$ term vanishes:

$$\mathbb{V}(X_k) = \mathbb{V}(X_k - X + X)$$
$$= \mathbb{V}(X_k - X) + \mathbb{V}(X) + \mathbb{E}((X_k - X)X)$$
$$= \mathbb{V}(X_k - X) + \mathbb{V}(X). \qquad (5)$$

Thus the exchangeability provides the justification for decomposing the total variance of an ensemble prediction system as a sum of the variance of the ensemble mean and ensemble dispersion about this mean (see also Siegert *et al.,* 2016, for a related discussion). This decomposition, which many readers will be familiar with, is used and implied in many of the references we refer to, including those that originally defined the RPC.

We may now invoke Pythagoras' theorem as before and conclude that, because the triangle formed by $\sqrt{\mathbb{V}(X)}$, $\sqrt{\mathbb{V}(X_k)}$, and $\sqrt{\mathbb{V}(X - X_k)}$ is right-angled, the correlation between X and X_k is given by ρ_σ, as claimed.

The geometric situation is shown in Figure 3; the solid triangle has the same interpretation as in Figure 1, formed by the square roots of $\mathbb{V}(X), \mathbb{V}(Y)$, and $\mathbb{V}(X - Y)$. The triangle with the dashed lines, however, is formed by the square roots of $\mathbb{V}(X), \mathbb{V}(X_k)$, and $\mathbb{V}(X - X_k)$, that is, the ensemble member replaces the verification. The cathetes opposite the angle α (respectively $\overline{\alpha}$) have the same length in both triangles, but only the latter triangle is a right-angled one. Moreover, we see from this construction that if the solid-line triangle is obtuse then the dashed triangle exhibits an angle $\overline{\alpha}$ between $\mathbb{V}(X)$ and $\mathbb{V}(X_k)$ which is larger than α. Therefore, if the solid-line triangle is obtuse, the correlation between X and X_k is smaller than that between X and Y. In other words, when $\overline{\alpha}$ is greater than α, an anomalous signal-to-noise ratio is obtained.

The term "paradox" has presumably been used in the literature due to an implicit (but generally unwarranted) assumption that the ensemble mean forecast X is linearly calibrated, or that the solid triangle in Figure 3 should also be right-angled (and therefore identical to the dashed triangle). In this case, $\overline{\alpha}$ is equal to α and the correlation of the ensemble mean with each ensemble member is identical to the correlation between the ensemble mean and the verification. In other words, this would imply RPC = 1. Effectively the same apparent paradox, yet in a different guise, would arise from the assumption that Pythagoras'
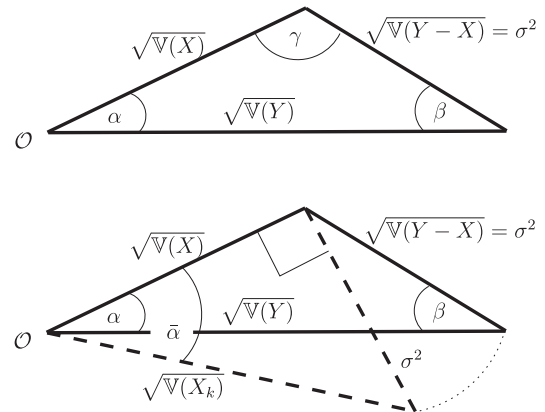

FIGURE 3 The upper panel shows a triangle corresponding to ensemble mean and verification (as in Figure 1, solid lines), the only difference being that the angle γ is now obtuse. The lower panel shows the same triangle overlaid with another triangle (dashed lines) in which the verification Y is replaced with an ensemble member X_k. Both triangles share the cathete of length $\sqrt{\mathbb{V}(X)}$. Furthermore, the cathete opposite the angles α (respectively $\overline{\alpha}$) has the same length in both triangles, but only the latter triangle is a right-angled one. As a consequence $\overline{\alpha} \geq \alpha$, resulting in an anomalous signal-to-noise ratio.

theorem applies to the verification in the same way that it does to the ensemble members. In reality, however, although

$$\mathbb{V}(X_k) = \sigma^2 + \mathbb{V}(X)$$

is always correct (this is Equation 5), the corresponding relation (Equation 2) for the verification *only* holds if the ensemble mean forecast X is linearly calibrated; in general we have

$$\mathbb{V}(Y) \neq \sigma^2 + \mathbb{V}(X).$$

Again in different words, the error between the ensemble mean and the ensemble members is always uncorrelated with the ensemble mean and can thus be regarded as noise; yet the error between the ensemble mean and the verification (whether or not we assume it to have the same magnitude as the error between the ensemble mean and the ensemble members) is in general *not* uncorrelated with the ensemble mean and thus cannot be regarded as noise. Therefore, the correlation between ensemble mean and ensemble members will in general not be equal to the correlation between ensemble mean and verification. The possibility of such a situation was first noted by Kumar (2009).

The ratio of predictable components RPC_σ ceases to be a useful diagnostic if the ensemble mean error $\mathbb{V}(Y - X)$ is very different from the ensemble spread σ^2, that is, if the ensemble is no longer marginally calibrated. In that situation, the ensemble does not contain any statistical

information about the verification beyond the ensemble mean, and consequently the correlation between ensemble mean and ensemble members cannot be expected to bear any relation to the correlation between ensemble mean and verification. Of course, the value of RPC makes no reference to σ^2 (as opposed to RPC_σ) and may therefore be used as a diagnostic even if the ensemble is not marginally calibrated. It must be kept in mind, however, that RPC is merely a diagnostic of the ensemble *mean* and its correlation with respect to the verification. Therefore RPC does not permit any conclusions about correlations between ensemble mean and ensemble members, in contrast to RPC_σ.

FIGURE 4   Ensemble mean X and dispersion σ^2 give rise to the dashed triangle. If $\mathbb{V}(X - Y) = \sigma^2$, the vertex $\mathcal{P}$ lies on the circle indicated. The signal-to-noise ratio will be normal if $\mathcal{P}$ lies on the solid section of the circle and anomalous if $\mathcal{P}$ lies on the dashed section. There are two dashed sections corresponding to upper and lower bounds in Equation (6), unless $\alpha \geq \pi/4$, in which case the smaller section disappears.

3.2 | Why an anomalous SNR is not a paradox

In Weisheimer *et al.* (2019) it is hypothesised that the paradox emerges due to statistical problems in estimating correlation-based statistics such as the ratio of predictable components. When one examines the RPC as above (or in other prior studies), it is clear that there is large potential for error in this estimate, as it is a ratio of two quantities containing further ratios of sample variances estimated from small samples with typically no more than 30 or 40 members. In other words, the RPC might in fact not be significantly larger (in a statistical sense) than one. Nonetheless, the ubiquitous nature of low signal-to-noise ratios in many models and model experiments and on different timescales perhaps suggests a different explanation.

Indeed, by considering the geometric relationships between correlation, RPC, and variance, we will now see that both normal and anomalous signal-to-noise ratios are possible, depending on the precise shape of the triangle in Figure 1. First we note that we may regard $\mathbb{V}(Y)$ as fixed and given by the problem statement and the verifications, while both $\mathbb{V}(X)$ and α are the variables that determine the triangle in Figure 1. (In fact we could assume $\sqrt{\mathbb{V}(Y)} = 1$ by dividing X and Y by $\sqrt{\mathbb{V}(Y)}$, as this would not change ρ or ρ_f, but have decided not to do so in this section for the sake of clarity.) We claim that the signal-to-noise ratio is normal if and only if

$$\rho \leq \sqrt{\frac{\mathbb{V}(X)}{\mathbb{V}(Y)}} \leq \frac{\rho}{2\rho^2 - 1}, \tag{6}$$

where the upper limit is interpreted as ∞ if $\rho^2 \leq 1/2$, which corresponds to $\alpha \geq \pi/4$. The claim will be proved at the end of this section.

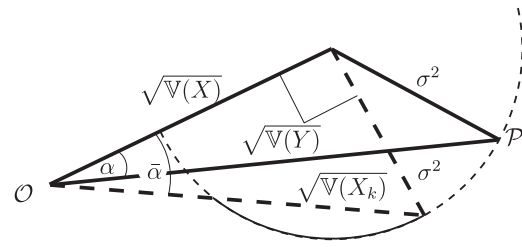The criterion in Equation (6) can be translated to the criterion in Christiansen *et al.* (2022), if the right identifications are made. The model introduced in Christiansen *et al.* (2022) uses X and O for the signal component of forecasts and observations, respectively, and assumes these are related through $X = \lambda O$. The signal O, however, must be defined in context with the forecast X, namely as the part of the observation that can be "explained" through the forecast X, while the "unexplained" noise $Y - O$ in the observation is then uncorrelated with X (see Siegert *et al.*, 2016, end of section 2a for a discussion of this point). This requires that $\lambda = \mathbb{V}(X)/\mathbb{E}(YX)$ (a value of $\lambda = 1$ would mean that X is linearly calibrated in the sense of the present work). With λ set in this way, the criterion in Christiansen *et al.* (2022) translates into the criterion in Equation (6) (both upper and lower bound).

Figure 4 illustrates the criterion in Equation (6) further. Given an ensemble forecast with mean X and dispersion σ^2, then, for any verification Y with the property that $\mathbb{V}(X - Y) = \sigma^2$, the solid triangle formed by $\mathbb{V}(X)$, $\mathbb{V}(Y)$, and $\mathbb{V}(X - Y)$ will have the vertex $\mathcal{P}$ lying on the circle indicated with radius σ^2. The signal-to-noise ratio will be normal if $\mathcal{P}$ lies on the solid section of the circle, while it will be anomalous if it lies on the dashed section. It is seen that there are two dashed sections, with the larger one corresponding to obtuse triangles and the lower bound in Equation (6), while the smaller section corresponds to very acute triangles and the upper bound in Equation (6). If $\alpha \geq \pi/4$, the latter section does not exist.

We stress that the RPC is always anomalous if the angle γ in the triangle in Figure 1 is obtuse, since this corresponds to the correlation $\rho = \cos(\alpha)$ being larger than $\cos \overline{\alpha} = \sqrt{\mathbb{V}(X)/\mathbb{V}(Y)}$, as is easily seen using elementary geometry and should also be clear from Figure 3. This seems to be the prevalent situation in climate forecasts, where forecasts exhibit significant correlations ρ with

the verifications but the variance of the ensemble mean $\mathbb{V}(X)$ is much smaller than the variance of the verifications $\mathbb{V}(Y)$ (as noted in Scaife *et al.,* 2014; Scaife and Smith, 2018). In fact, the correlation does not need to be that large. Assuming that $\mathbb{V}(X)/\mathbb{V}(Y)$ is of order 0.1 (already an overestimate for seven out of the eight forecasting systems we will consider in Section 4), the RPC becomes anomalous as soon as the correlation exceeds $\rho = 0.316$ (which corresponds to an angle α smaller than about 72°).

Another observation, again relevant in the seasonal forecasting context, pertains to the sensitivity of RPC, the ratio of predictable components, with respect to changes in the correlation ρ. We consider a situation that is calibrated linearly initially and vary the correlation ρ but keep $\mathbb{V}(Y)$ and β constant (see Figure 3 for definition of β). We find (using elementary calculations) that

$$\frac{d(\text{RPC})}{d\rho} = \frac{1}{\rho}(1 - \rho^2). \tag{7}$$

Clearly, the smaller the correlation, the more sensitive RPC will be to a small increase or decrease in the correlation, which can easily occur in small samples with 30 or 40 years of data.

A similar calculation also explains the phenomenon (reported in several papers, e.g., Scaife *et al.,* 2014) of an anomalous signal-to-noise ratio despite $\mathbb{V}(X_k)$ and $\mathbb{V}(Y)$ being *approximately* equal. Note first that if $\mathbb{V}(X_k) = \mathbb{V}(Y)$ (and $\sigma^2 = \mathbb{E}[(X - Y)^2]$ as we always assume), then RPC = 1 and there is no signal-to-noise paradox. Despite this, in situations of small correlations (as in climate forecasts), minor differences between $\mathbb{V}(X_k)$ and $\mathbb{V}(Y)$ can lead to strongly anomalous signal-to-noise ratios. To see this, we again consider a situation that is calibrated linearly initially (this implies $\mathbb{V}(X_k) = \mathbb{V}(Y)$ and RPC = 1), but now we vary $\mathbb{V}(Y)$ and keep $\mathbb{V}(X)$ and σ^2 constant. Again, using elementary calculations we find

$$\frac{d(\text{RPC})}{d(\mathbb{V}(Y))} = \frac{1}{2\sigma^2}\left(\frac{1 - \rho^2}{\rho}\right)^2. \tag{8}$$

This relationship has an equivalent form in terms of logarithmic derivatives, thus describing *relative* changes of RPC due to relative changes of $\mathbb{V}(Y)$:

$$\frac{d[\log(\text{RPC})]}{d[\log \mathbb{V}(Y)]} = \frac{1}{2}\left(\frac{1}{\rho^2} - 1\right). \tag{9}$$

Equation (9) again shows that, for small correlations ρ^2, minor violations of the identity $\mathbb{V}(X_k) = \mathbb{V}(Y)$ can have a large effect on RPC.

3.3 | SNR is normal if and only if Equation (6) holds

We will now demonstrate our claim that the signal-to-noise ratio is normal if and only if the relation in Equation (6) holds. We note that RPC = 1 if and only if

$$\rho^2 = \frac{\mathbb{V}(X)}{\mathbb{V}(X) + \mathbb{V}(Y - X)}, \tag{10}$$

or, after multiplying with $\mathbb{V}(X) + \mathbb{V}(Y - X)$ and subtracting $\rho^2\mathbb{V}(X)$, we find

$$\rho^2\mathbb{V}(Y - X) = (1 - \rho^2)\mathbb{V}(X), \tag{11}$$

After expanding the left-hand side, this becomes

$$\rho^2(\mathbb{V}(Y) + \mathbb{V}(X) - 2\rho\sqrt{\mathbb{V}(Y)\mathbb{V}(X)}) = (1 - \rho^2)\mathbb{V}(X). \tag{12}$$

Dividing by $\mathbb{V}(Y)$ and calling $\xi := \sqrt{\mathbb{V}(X)/\mathbb{V}(Y)}$, we finally arrive at

$$\rho^2 + (2\rho^2 - 1)\xi^2 - 2\rho^3\xi = 0. \tag{13}$$

The roots of the left-hand side are $\xi_1 = \rho$ and $\xi_2 = \rho/(2\rho^2 - 1)$. If $\rho^2 > 1/\sqrt{2} \cong 0.71$, then RPC < 1 for $\xi_1 < \xi < \xi_2$. If $\rho^2 < 1/\sqrt{2}$, then the second root ξ_2 becomes negative and we get RPC < 1 for $\xi_1 < \xi$. These two statements together prove our claim around the relation in Equation (6). As a final note, we stress that, in any event, whenever the lower bound in Equation (6) fails to hold and we have $\sqrt{\mathbb{V}(X)/\mathbb{V}(Y)} < \rho$, then the solid triangle in Figure 3 is obtuse and RPC > 1, that is, an anomalous signal-to-noise ratio. Underlying the analysis are two important assumptions. Firstly, the statistical quantities such as $\mathbb{V}(X), \mathbb{V}(Y), \mathbb{V}(Y - X)$, as well as the average ensemble spread σ^2 and the correlation ρ, are assumed not to depend on time. Secondly, we assume the ensemble members to be exchangeable, or, in other words, indistinguishable with regards to their statistical properties (Bröcker and Kantz, 2011). We stress that assumptions made in previous work are at least as strong, and there is no need here for a more specific model (as in Siegert *et al.,* 2016; Charlton-Perez *et al.,* 2019; Christiansen *et al.,* 2022, for example).

4 | NUMERICAL SIMULATIONS: NORMAL AND ANOMALOUS RPC OF ECMWF SEASONAL HINDCASTS

In this section we consider data from several different seasonal hindcast experiments. All diagnostics are for a

North Atlantic Oscillation (NAO) index used routinely at ECMWF, namely the first EOF of Z500, averaged over the three months of December–January–February. The forecasts are taken from four different hindcast data sets listed below; all forecasting systems use 25 ensemble members, which are initialised on each November 1 during the hindcast period.

SEAS5   is ECMWF's operational seasonal forecasting system (Johnson *et al.,* 2019), run in spectral atmospheric resolution Tco319 (approx 40 km) with 91 vertical levels and the 1/4 degree horizontal resolution ocean configuration with 75 vertical levels (ORCA025L75), with forecasts available from 1981 onwards.

ASF   is the Atmospheric Seasonal Forecasts of the 20th Century (ASF-20C) experiment (Weisheimer *et al.,* 2017). The configuration is atmosphere-only with prescribed observed SSTs run in resolution T255L91 (approx. 80 km horizontally) and a slightly older model version, with forecasts available from 1901–2010. Note that, since these forecasts require future SST data as boundary conditions at the start of the forecast, they should be considered idealised forecast experiments.

CSF   is the Coupled Seasonal Forecasts of the 20th Century (CSF-20C) experiment (Weisheimer *et al.,* 2020). The configuration is similar to ASF but run with a coupled model that uses a resolution of 1 degree and 42 vertical levels in the ocean, with forecasts available from 1901–2010.

The experiments compare each of these hindcasts with two sets of verifications. The first set of verifications is taken from ERA-5, which is ECMWF's latest atmosphere-only reanalysis (Hersbach et al., 2020). (We stress that reanalyses are not direct meteorological observations, but rather meteorological fields obtained by assimilation of historical observations into a climate model.) The second set is taken from CERA-20C, which is ECMWF's first coupled reanalysis of the 20th Century (which was also used to initialise the CSF experiment Laloyaux *et al.,* 2018).

The common hindcast period between all three experiments is 1981–2009, and results are shown in Table 1 for the ERA-5 verifications and in Table 2 for the CERA-20C verifications. In the context of the analysis presented in Section 2, X is the mean of an ensemble with members $X_1, \ldots, X_K$, while Y is the verification. Only $\mathbb{V}(Y)$ (which is equal to one, due to data

normalisation), $\mathbb{V}(X)$, $\mathbb{V}(Y-X)$, and the ensemble dispersion $\sigma^2 = \mathbb{V}(X-X_k)$ are relevant for the analysis in the present section. It is evident from visual inspection of Tables 1 and 2 that the seasonal hindcast statistics for ERA-5 verifications and CERA-20C verifications are very similar. In the following, we will therefore focus on the results for CERA-20C verifications in this hindcast period.

Table 3 shows results for the two experiments ASF and CSF, but for an earlier hindcast period (1926–1954). In this period, the skill and signal-to-noise behaviour are different from those in the more recent common period. The verification data are CERA-20C (because ERA5 data are not available for these early years).

All quantities in rows 4 and below in these tables are functions of those in the first three rows. The quantities $\mathbb{V}(X)$ and $\mathbb{V}(Y-X)$ completely determine the triangle and hence also allow us to compute ρ. It is seen that, for all forecasting systems over the common hindcast period 1981–2009, we have $\mathbb{V}(X)/\mathbb{V}(Y) = \mathbb{V}(X) < \rho^2$ and hence these should, according to our analysis, exhibit an anomalous signal-to-noise ratio, which they do as RPC > 1 for these systems.

The triangles are shown in Figure 5. For all forecasting systems over the recent common hindcast period 1981–2009 (Figure 5, lower panel) it is readily apparent that the angle γ (at the top corner of the triangles) is obtuse, and hence that these forecasting systems exhibit an anomalous signal-to-noise ratio. For the early hindcast period from 1926–1954, however, γ is an acute angle, as can be seen from Table 3 and also from Figure 5 (upper panel, probably with the help of a geometry set square). In any event, since $\rho^2 < 1/2$ for both forecasting systems considered, the upper limit in Equation (6) is to be taken as infinity, so, in order to check whether these systems exhibit a normal signal-to-noise ratio or not, $\mathbb{V}(X)$ has to be compared with ρ^2.

TABLE 1   Seasonal hindcast statistics against ERA-5 for the common verification period 1981–2009. Forecasts and verifications have been normalised so that $\mathbb{V}(Y) = 1$.

	SEAS5	ASF	CSF
$\mathbb{V}(X)$	0.065	0.081	0.070
$\mathbb{V}(X-Y)$	0.894	0.776	0.814
σ^2	0.734	0.661	0.698
ρ^2	0.112	0.287	0.234
ρ	0.335	0.536	0.484
RPC	1.288	1.743	1.719
RPC_σ	1.176	1.622	1.602

TABLE 2 Seasonal hindcast statistics against CERA-20C for the common verification period 1981–2009. Forecasts and verifications have been normalised so that $\mathbb{V}(Y) = 1$. Note that normalisations in Tables 1 and 2 are numerically slightly different, as the verifying data sets are different. This is the reason for the (very slightly) different numerical values of $\mathbb{V}(X)$ and σ^2 between these tables, which in theory should be the same.

	SEAS5	ASF	CSF
$\mathbb{V}(X)$	0.066	0.082	0.070
$\mathbb{V}(X - Y)$	0.895	0.771	0.813
σ^2	0.737	0.664	0.702
ρ^2	0.111	0.295	0.236
ρ	0.333	0.543	0.486
RPC	1.270	1.751	1.725
RPC_σ	1.161	1.638	1.613

TABLE 3 Seasonal hindcast statistics against CERA-20C for the verification period 1926–1954. Forecasts and verifications have been normalised so that $\mathbb{V}(Y) = 1$.

	ASF	CSF
$\mathbb{V}(X)$	0.132	0.076
$\mathbb{V}(X - Y)$	0.966	0.985
σ^2	0.938	0.973
ρ^2	0.050	0.027
ρ	0.224	0.164
RPC	0.666	0.617
RPC_σ	0.657	0.613

We have also shown the ensemble dispersion $\sigma^2 = \mathbb{V}(X - X_k)$. With this, in Section 3 we defined the quantity

$$\mathrm{RPC}_\sigma := \rho / \sqrt{\frac{\mathbb{V}(X)}{\mathbb{V}(X) + \sigma^2}},$$

which (as we recall) should be seen as an alternative to RPC where the error $\mathbb{V}(X - Y)$ is replaced with the ensemble dispersion σ^2. Clearly, RPC and RPC_σ should coincide if the ensemble dispersion describes the magnitude of the ensemble mean error correctly. This is not quite the case for the most recent verification periods in Tables 1 and 2, as the ensembles are all underdispersive for these verification periods, with the result that $RPC > RPC_\sigma$. Since all these forecasts exhibit an anomalous signal-to-noise ratio with respect to RPC_σ, the signal-to-noise ratio is "even more anomoalous" with respect to RPC. In other words, both quantities give the same conclusions.

For the early verification periods in Table 3, the ensemble dispersion matches the magnitude of the ensemble
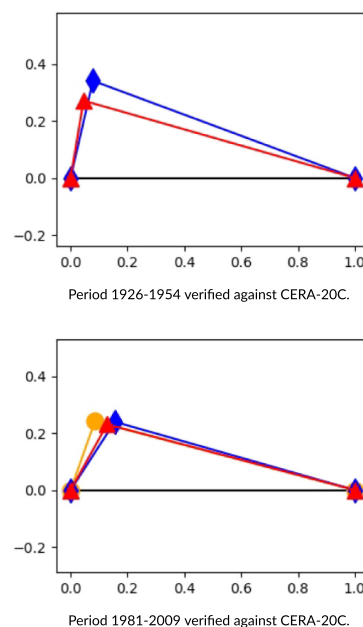

Period 1926-1954 verified against CERA-20C.

Period 1981-2009 verified against CERA-20C.

FIGURE 5 Triangles as in Figure 1 for hindcast sets SEAS5 (●), ASF (◆), and CSF (▲). As in Figure 1, the horizontal line represents $\sqrt{\mathbb{V}(Y)}$, the shorter cathete on the left is $\sqrt{\mathbb{V}(X)}$, and the longer one on the right is $\sqrt{\mathbb{V}(X - Y)}$. Upper panel: 1926–1954 period verified against CERA-20C; lower panel: 1981–2009 period verified against CERA-20C. Results for evaluation against ERA-5 during this period are virtually the same. [Colour figure can be viewed at wileyonlinelibrary.com]

mean error to a higher degree, although the ensembles are still a bit underdispersive. Hence, for these verification periods also, we find that $RPC > RPC_\sigma$, although the difference is much smaller. In contrast to the experiments for the later verification period, the forecasts now exhibit a *normal* signal-to-noise ratio with respect to RPC, and hence also with respect to RPC_σ. Again, both quantities give the same conclusions.

5 | CONCLUSIONS AND FUTURE WORK

In ensemble forecasting, the ratio of predictable components, or RPC, is given by the correlation of the ensemble mean with the verification, divided by the correlation of the ensemble mean with individual members. For a perfect ensemble in which ensemble members are statistically indistinguishable from the verification, the RPC should be equal to one. Many cases have been reported of ensemble forecasting systems exhibiting RPC values significantly different from one; cases with RPC > 1 have received particular attention, as this implies that the ensemble mean is better at forecasting the verification than the ensemble members (in the sense that it is stronger correlated with the former than with the latter).

In this contribution, the RPC has been analysed from a statistical perspective. It has been shown that the properties of the RPC are easily understood in terms of the triangle formed by the (square root of the) variances of the ensemble mean and verification, as well as the ensemble spread. An anomalous signal-to-noise ratio is a consequence of the relative magnitudes of these three quantities, and in particular occurs if the mentioned triangle is obtuse. This, as we have seen, happens as soon as the correlation ρ between the ensemble mean and verification exceeds the ratio $\sqrt{\mathbb{V}(X)/\mathbb{V}(Y)}$ of the variance of the ensemble mean to the observation variance. A ratio $\mathbb{V}(X)/\mathbb{V}(Y)$ of order 0.1, for instance, (a value not untypical for the seasonal hindcasts considered in the present article) implies that the RPC becomes anomalous as soon as the correlation exceeds $\rho = 0.316$; for $\mathbb{V}(X)/\mathbb{V}(Y) = 0.2$ the critical value is $\rho = 0.447$.

In medium-range forecasts, the signal-to-noise paradox is much less prevalent, presumably because the variance $\mathbb{V}(X)$ is much larger relative to $\mathbb{V}(Y)$. At the same time, the correlation between the ensemble mean and verification will also be larger, with the consequence that the RPC is more robust with respect to sampling variations (this follows from Equation 7).

This analysis also suggests that in certain other cases the RPC might be fairly sensitive to sampling variability of the quantities involved. To a degree, this is unavoidable, as any diagnostic quantity with discriminative power will necessarily exhibit a certain amount of sensitivity to sampling variability; however, this needs to be understood properly and controlled better. The RPC (or some other suitable quantity) might then be used as a statistic in a hypothesis test for the presence of an anomalous signal-to-noise ratio. This will be the subject of future work.

AUTHOR CONTRIBUTIONS

DATA AVAILABILITY STATEMENT

ERA-5: Reanalysis freely available via Copernicus C3S Climate Data Store (CDS, 2022).

See also Hersbach et al. (2020) for more information.

CERA-20C: ECMWF provides access to CERA–20C and other public data sets via a dedicated data portal (MARS, 2022). See also Laloyaux *et al.* (2018) for more information.

SEAS5: These seasonal forecasts are freely available via Copernicus C3S Climate Data Store (CDS, 2022).

ASF and CSF: These data are freely available through a dedicated online dissemination platform (Weisheimer *et al.,* 2020). A set of standard monthly mean atmospheric variables including temperature, precipitation, mean sea-level pressure, geopotential height, wind, and thermal and radiative fluxes have been provided as global gridded data in netCDF format.

ORCID

Jochen Bröcker https://orcid.org/0000-0002-0864-6530
Andrew J. Charlton–Perez https://orcid.org/0000-0001-8179-6220
Antje Weisheimer https://orcid.org/0000-0002-7231-6974

REFERENCES

Baker, L.H., Shaffrey, L.C., Sutton, R.T., Weisheimer, A. and Scaife, A.A. (2018) An intercomparison of skill and overconfidence/underconfidence of the wintertime North Atlantic Oscillation in multimodel seasonal forecasts. *Geophysical Research Letters*, 45(15), 7808–7817.

Bröcker, J. (2021) Testing the reliability of forecasting systems. *Journal of Applied Statistics*, 50(1), 106–130. https://doi.org/10.1080/02664763.2021.1981833.

Bröcker, J. and Kantz, H. (2011) The concept of exchangeability in ensemble forecasting. *Nonlinear Processes in Geophysics*, 18(1), 1–5, https://doi.org/10.5194/npg-18-1-2011.

Byrne, N.J., Shepherd, T.G. and Polichtchouk, I. (2019) Subseasonal-to-seasonal predictability of the southern hemisphere eddy-driven jet during austral spring and early summer. *Journal of Geophysical Research: Atmospheres*, 124(13), 6841–6855, https://doi.org/10.1029/2018JD030173.

C3S Climate Data Store (CDS). (2022) Copernicus Services, https://climate.copernicus.eu/climate-data-store.

Charlton-Perez, A.J., Bröcker, J., Stockdale, T.N. and Johnson, S. (2019) When and where do ECMWF seasonal forecast systems exhibit anomalously low signal-to-noise ratio? *Quarterly Journal of the Royal Meteorological Society*, 145, 3466–3478. https://doi.org/10.1002/qj.3631.

Christiansen, B., Yang, S. and Matte, D. (2022) The forced response and decadal predictability of the North Atlantic Oscillation: Nonstationary and fragile skills. *Journal of Climate*, 35(18), 5869–5882, https://doi.org/10.1175/JCLI-D-21-0807.1.

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L. and Robinson, N. (2014) Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, 41(15), 5620–5628.

Hardiman, S.C., Dunstone, N.J., Scaife, A.A., Smith, D.M., Comer, R., Nie, Y. and Ren, H.-L. (2022) Missing eddy feedback may explain weak signal-to-noise ratios in climate predictions. *npj Climate and Atmospheric Science*, 5(1), 57. https://doi.org/10.1038/s41612-022-00280-4.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, Joaquín, Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and Thépaut, J.-N. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049.

Johnson, S.J., Stockdale, T.N., Ferranti, L., Balmaseda, M.A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S.P.E., Mogensen, K., Zuo, H. and Monge–Sanz, B.M. (2019) SEAS5: The new ECMWF seasonal forecast system. *Geoscientific Model Development*, 12(3), 1087–1117, https://doi.org/10.5194/gmd-12-1087-2019.

Kumar, A. (2009) Finite samples and uncertainty estimates for skill measures for seasonal prediction. *Monthly Weather Review*, 137(8), 2622–2631, https://doi.org/10.1175/2009MWR2814.1.

Laloyaux, P., de Boisseson, E., Balmaseda, M., Bidlot, J.-R., Broennimann, S., Buizza, R., Dalhgren, P., Dee, D., Haimberger, L., Hersbach, H., Kosaka, Y., Martin, M., Poli, P., Rayner, N., Rustemeier, E. and Schepers, D. (2018) CERA–20C: A coupled reanalysis of the twentieth century. *Journal of Advances in Modeling Earth Systems*, 10(5), 1172–1195.

Leutbecher, M. and Palmer, T.N. (2008) Ensemble forecasting. *Journal of Computational Physics*, 227, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014.

Meteorological Archival and Retrieval System (MARS). (2022) European Centre for Medium Range Weather Forecasts, https://apps.ecmwf.int/datasets/data/cera20c/levtype=sfc/type=an/.

O'Reilly, C.H., Weisheimer, A., Woollings, T., Gray, L.J. and MacLeod, D. (2019) The importance of stratospheric initial conditions for winter North Atlantic Oscillation predictability and implications for the signal-to-noise paradox. *Quarterly Journal of the Royal Meteorological Society*, 145(718), 131–146, https://doi.org/10.1002/qj.3413.

Palmer, T.N. (2019) Stochastic weather and climate models. *Nature Reviews Physics*, 1(7), 463–471, https://doi.org/10.1038/s42254-019-0062-2.

Palmer, T.N., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M. and Smith, L.A. (2006) Ensemble prediction: A pedagogical perspective. In: *10th Workshop on Meteorological Operational Systems, 14 to 18 November 2005*. Shinfield Park, Reading: ECMWF, pp. 50–58, https://www.ecmwf.int/node/11525.

Rodgers, J.L. and Nicewander, W.A. (1988) Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1), 59–66.

Scaife, A.A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R.T., Dunstone, N., Eade, R., Fereday, D., Folland, C.K., Gordon, M.,

Hermanson, L., Knight, J.R., Lea, D.J., MacLachlan, C., Maidens, A., Martin, M., Peterson, A.K., Smith, D., Vellinga, M., Wallace, E., Waters, J. and Williams, A. (2014) Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, 41(7), 2514–2519.

Scaife, A.A. and Smith, D. (2018) A signal-to-noise paradox in climate science. *npj Climate and Atmospheric Science*, 1(1), 28, https://doi.org/10.1038/s41612-018-0038-4.

Seviour, W.J.M., Hardiman, S.C., Gray, L.J., Butchart, N., MacLachlan, C. and Scaife, A.A. (2014) Skillful seasonal prediction of the southern annular mode and Antarctic Ozone. *Journal of Climate*, 27(19), 7462–7474, https://doi.org/10.1175/JCLI-D-14-00264.1.

Siegert, S., Stephenson, D.B., Sansom, P.G., Scaife, A.A., Eade, R. and Arribas, A. (2016) A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability? *Journal of Climate*, 29(3), 995–1012.

Smith, D.M., Scaife, A.A., Eade, R., Athanasiadis, P., Bellucci, A., Bethke, I., Bilbao, R., Borchert, L.F., Caron, L.P., Counillon, F., Danabasoglu, G., Delworth, T., Doblas-Reyes, F.J., Dunstone, N.J., Estella-Perez, V., Flavoni, S., Hermanson, L., Keenlyside, N., Kharin, V., Kimoto, M., Merryfield, W.J., Mignot, J., Mochizuki, T., Modali, K., Monerie, P.A., Müller, W.A., Nicolí, D., Ortega, P., Pankatz, K., Pohlmann, H., Robson, J., Ruggieri, P., Sospedra-Alfonso, R., Swingedouw, D., Wang, Y., Wild, S., Yeager, S., Yang, X. and Zhang, L. (2020) North Atlantic climate far more predictable than models imply. *Nature*, 583(7818), 796–800.

Stockdale, T.N., Molteni, F. and Ferranti, L. (2015) Atmospheric initial conditions and the predictability of the arctic oscillation. *Geophysical Research Letters*, 42(4), 1173–1179.

Strommen, K. and Palmer, T.N. (2019) Signal and noise in regime systems: A hypothesis on the predictability of the North Atlantic Oscillation. *Quarterly Journal of the Royal Meteorological Society*, 145(718), 147–163, https://doi.org/10.1002/qj.3414.

Weigel, A.P. (2011) Verification of ensemble forecasts. In: Jolliffe, I.T. and Stephenson, D.B. (Eds.) *Forecast Verification; A Practicioner's Guide in Athmospheric Science*, 2nd edition. Chichester: John Wiley & Sons, Ltd, pp. 141–166.

Weisheimer, A., Befort, D.J., MacLeod, D., Palmer, T.N., O'Reilly, C. and Strømmen, K. (2020) Seasonal forecasts of the twentieth century. *Bulletin of the American Meteorological Society*, 101(8), E1413–E1426, https://doi.org/10.1175/BAMS-D-19-0019.1.

Weisheimer, A., Decremer, D., MacLeod, D., O'Reilly, C., Stockdale, T.N., Johnson, S. and Palmer, T.N. (2019) How confident are predictability estimates of the winter North Atlantic Oscillation? *Quarterly Journal of the Royal Meteorological Society*, 145(S1), 140–159, https://doi.org/10.1002/qj.3446.

Weisheimer, A. and Palmer, T.N. (2014) On the reliability of seasonal climate forecasts. *Journal of The Royal Society Interface*, 11(96), 20131162, https://doi.org/10.1098/rsif.2013.1162.

Weisheimer, A., Palmer, T.N. and Doblas-Reyes, F.J. (2011) Assessment of representations of model uncertainty in monthly and seasonal forecast ensembles. *Geophysical Research Letters*, 38(16).

Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D.A. and Palmer, T.N. (2017) Atmospheric seasonal forecasts of the twentieth century: Multi-decadal variability in predictive skill of the winter North Atlantic Oscillation (NAO) and their potential value for extreme event attribution. *Quarterly Journal of the Royal Meteorological Society*, 143(703), 917–926, https://doi.org/10.1002/qj.2976.

Zhang, W. and Kirtman, B. (2019) Understanding the signal-to-noise paradox with a simple Markov model. *Geophysical Research Letters*, 46(22), 13308–13317, https://doi.org/10.1029/2019GL085159.

Zhang, W., Kirtman, B., Siqueira, L., Clement, A. and Xia, J. (2021) Understanding the signal-to-noise paradox in decadal climate predictability from CMIP5 and an eddying global coupled model. *Climate Dynamics*, 56(9), 2895–2913, https://doi.org/10.1007/s00382-020-05621-8.