Check for updates

Geophysical Research Letters

RESEARCH LETTER
10.1029/2023GL103710

On the Relationship Between Reliability Diagrams and the "Signal-To-Noise Paradox"

Kristian Strommen[1], **Molly MacRae**[2], **and Hannah Christensen**[1]

[1]Department of Physics, University of Oxford, Oxford, UK, [2]Centre for Environmental Data Analysis, Oxfordshire, UK

Key Points:

- The exact relationship between the "signal-to-noise paradox" and standard measures of forecast reliability has remained unclear
- It is shown that the "paradox" is equivalent to the slope of reliability diagrams exceeding 1, assuming linearity/Gaussianity
- The value of reliability diagrams as a diagnostic tool for understanding the "paradox" better is discussed

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

K. Strommen,
kristian.strommen@physics.ox.ac.uk

Citation:

Strommen, K., MacRae, M., & Christensen, H. (2023). On the relationship between reliability diagrams and the "signal-to-noise paradox". *Geophysical Research Letters*, *50*, e2023GL103710. https://doi.org/10.1029/2023GL103710

Received 15 MAR 2023
Accepted 1 JUL 2023

© 2023 The Authors.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Abstract The "signal-to-noise paradox" for seasonal forecasts of the winter North Atlantic Oscillation (NAO) is often described as an "underconfident" forecast and measured using the ratio-of-predictable components (RPCs) metric. However, comparison of RPC with other measures of forecast confidence, such as spread-error ratios, can give conflicting impressions, challenging this informal description. We show, using a linear statistical model, that the "paradox" is equivalent to a situation where the reliability diagram of any percentile forecast has a slope exceeding 1. The relationship with spread-error ratios is shown to be far less direct. We furthermore compute reliability diagrams of winter NAO forecasts using seasonal hindcasts from the European Centre for Medium-range Weather Forecasts and the UK Meteorological Office. While these broadly exhibit slopes exceeding 1, there is evidence of asymmetry between upper and lower terciles, indicating a potential violation of linearity/Gaussianity. The limitations and benefits of reliability diagrams as a diagnostic tool are discussed.

Plain Language Summary The North Atlantic Oscillation (NAO) is an atmospheric phenomenon which can be understood as summarizing large-scale winter conditions across western Europe. Long-range forecasts of the NAO have been shown to be skillful, but also to suffer from a so-called "signal-to-noise paradox," which roughly says that the real world appears to be more predictable than the forecasts think it is. However, interpreting the exact meaning of this "paradox" has proved challenging. We help bring some clarity by showing that one can interpret the "paradox" as a case of a probabilistically underconfident forecast, namely a forecast which tends to underestimate the likelihood of high magnitude NAO events.

1. Introduction

It is now well established that weather forecasting models are able to generate skillful seasonal forecasts of the winter North Atlantic Oscillation (NAO) (Athanasiadis et al., 2017; Dunstone et al., 2016; Eade et al., 2014; Scaife et al., 2014). However, these forecasts also suffer from a curious phenomenon dubbed a "signal-to-noise paradox" (Dunstone et al., 2016; Eade et al., 2014). An overview and discussion of this phenomenon is given by Scaife and Smith (2018), who also explain why understanding, and ultimately eliminating the "paradox" from forecasts is a problem of great practical importance.

The "paradox" is commonly measured using a correlation based metric referred to as the "ratio-of-predictable components" (RPC), as introduced in Eade et al. (2014). A forecast is said to exhibit a "signal-to-noise paradox" when RPC > 1, which corresponds to a situation where the ensemble mean is a better predictor of the real world than of individual ensemble members (see Section 2.4). However, interpreting this situation and understanding how RPC relates to other skill metrics has proved challenging (Bröcker et al., 2023). Indeed, the choice of the word "paradox" suggests that this phenomenon is often viewed as strange and unintuitive by the weather forecasting community.

Eade et al. (2014) interpreted the "paradox" as an "underconfident" forecast by computing reliability diagrams of mean sea-level pressure gridpoint forecasts and observing that forecasts with RPC > 1 produced reliability diagrams with a slope exceeding 1. Reliability diagrams (Murphy, 1973) have been emphasized as an important tool to study seasonal forecasts in a probabilistic manner (Weisheimer & Palmer, 2014). They offer an intuitive and easy-to-interpret measure of confidence: a forecast of a binary event E can be thought of as "underconfident" if, in situations where the forecast probability P_f of E occurring is low, event E actually occurs in the real world with a frequency greater than P_f. In other words, the forecast model underestimates the true probability of E occurring. Similarly, an "overconfident" forecast would be one which overestimates the true probability.

19448007, 2023, 14, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL103710 by Oxford University, Wiley Online Library on [03/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

However, in Strommen and Palmer (2019) it was shown that spread-error ratios, another metric widely used to measure forecast confidence (Johnson & Bowler, 2009), do not always give the same qualitative conclusion as the RPC. In fact, one can easily construct statistical forecast models that are "underconfident" with respect to RPC but "overconfident" with respect to spread-error (Strommen & Palmer, 2019), and there is evidence suggesting such a mismatch actually occurs in the case of winter NAO forecasts (Weisheimer et al., 2019). The goal of the present paper is therefore to address the following questions.

1. What exactly is the relationship between RPC, reliability diagrams and spread-error ratios?
2. Can reliability diagrams of winter NAO forecasts shed further light on the "paradox"?

To address these questions, we will make use of two types of data. First, we use artificially generated data based on the linear statistical model of Siegert et al. (2016). This will allow us to assess reliability and its relation to RPC in an idealized situation where, in particular, the sample size can be made large enough to minimize noise. In fact, the explicit nature of the statistical model allows for a theoretical comparison between reliability, RPC and spread-error. Second, we will compute and assess the NAO forecast reliability of two world-leading forecast models: the UK Met Office model (UKMO) (Dunstone et al., 2016) and the European Centre for Medium-range Weather Forecasts (ECMWF) model (Weisheimer et al., 2017).

2. Data and Methods

2.1. Data

The UKMO hindcast data used is the 40-member "DePreSys3" ensemble, based on the HadGEM3-GC2 version of Met Office Unified Model described in Dunstone et al. (2016). The data set consists of 35 ensemble forecasts initialized on November 1st for every year between 1980 and 2015. The forecast model includes interactive atmosphere-ocean coupling and has a nominal atmospheric resolution of 0.83° longitude by 0.55° latitude with 85 vertical levels. The nominal ocean resolution is 0.25°. The ensemble mean NAO correlations attained are approximately 0.6.

The ECMWF data used is the 51-member ensemble "ASF20C," based on a version of the Integrated Forecast System closely related to System 4 forecast system (Molteni et al., 2011). ASF20C consists of 110 ensemble forecasts initialized on November 1st for every year between 1900 and 2010. The horizontal spectral resolution of the model of T255 corresponds to a grid length of approximately 80 km with 91 vertical levels. ASF20C is uncoupled, and uses prescribed SSTs from the ERA20C reanalysis (Poli et al., 2016). Further details can be found in Weisheimer et al. (2017). The ensemble mean NAO correlations attained over the period 1980–2010 are approximately 0.5.

We use ERA20C as our observational "truth" in order to allow for a comparison with ASF20C across the whole 20th century. In order to have a clean comparison with the UKMO data, we will consider the 31 winters covering 1980–2010 ($N = 31$), as well as the full period 1900–2010 ($N = 109$). We will only consider DJF mean quantities.

2.2. Definition of the NAO Index

The NAO timeseries for ERA20C and all ECMWF/UKMO ensemble members are defined as in Dunstone et al. (2016), namely as the difference in DJF-averaged mean sea-level pressure anomalies between Iceland (63–70N, 25–16W) and the Azores (36–40N, 28–20W). The timeseries are normalized to have mean 0 and standard deviation 1.

2.3. Reliability Diagrams and Definition of the Forecast Events

There are a wealth of resources concerning reliability diagrams (Bröcker & Smith, 2007; Murphy & Winkler, 1977; Weisheimer & Palmer, 2014). For completeness we provide the basic definitions.

Suppose we have an ensemble forecast of a binary event E at times t consisting of ensemble members $x_{t,k}$, $k = 1, \ldots R$, where R is the ensemble size. Let y_t be the timeseries of E as observed in the real world. At time t, the forecast probability P_f of E occurring is defined as the proportion of ensemble members for which E occurs. By considering all times t that share approximately the same forecast probability P_f, we can compute the proportion P_r of such times in which y_t registered an occurrence of E. A reliability diagram is simply a plot of P_f against P_r.

19448007, 2023, 14, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL103710 by Oxford University, Wiley Online Library on [03/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

A reliable forecast is one where $P_f = P_r$, which is guaranteed for a perfectly calibrated ensemble. In this case the reliability diagram coincides with the diagonal and the slope of a linear fit to the data will be 1. A forecast is said to be unreliable if the reliability diagram deviates from the diagonal. We therefore obtain a quantitative measure of the reliability of a forecast by estimating the slope of the reliability diagram. This measure is only sensible in cases where the relationship between P_f and P_r is approximately linear. Reliability diagrams computed using weather forecast data can often deviate strongly from linearity, but we will see that in our case the assumption of linearity is reasonable.

When applying this framework to winter NAO forecasts we follow standard conventions by defining two binary events in terms of the upper and lower tercile of the distribution. These events can jointly be thought of as assessing the reliability of forecasts predicting a notable deviation from neutral conditions. The forecast (observed) probability is computed with respect to the terciles of the forecast (observational) distribution, to avoid overpenalizing.

Probability/occurrence bins are defined for each decile (0%–10%, 10%–20%, etc.). When fitting a straight line to the raw scatter plot, the bins are weighted according to the number of samples they contain.

2.4. Statistical Testing and the RPC Metric

For significance tests, we use Monte Carlo resampling: generate 1,000 random samples, compute the relevant metric in all 1,000 cases, and use the resulting distribution to generate confidence intervals. With the SN-model (defined in the next section), random pairs of "observations" and "forecasts" are generated by taking random draws from the distributions of s, ϵ, and η and using the SN-model equations. When considering UKMO/ECMWF forecast data, random draws are generated by resampling years randomly with replacement to obtain shuffled timeseries of the same length as the original.

The RPC metric is defined by the formula

$$RPC = \frac{\sqrt{Corr(EnsMean, Obs)^2}}{\sqrt{\sigma_{sig}^2/\sigma_{tot}^2}}, \tag{1}$$

where *EnsMean* is the ensemble mean timeseries, *Obs* is the observational timeseries, σ_{sig}^2 is the ensemble mean variance, σ_{tot}^2 is the average variance of individual ensemble members, and the square root is always taken to be positive. Eade et al. (2014) motivate this metric, and its name, by noting that if the forecast has skill, then the RPC is a lower bound approximation to the ratio $PC(Obs)/PC(Mod)$, where the numerator (denominator) is the square root of the proportion of variance that is predictable in the real world (the forecast world). It can be shown (Strommen & Palmer, 2019) that

$$RPC \approx \frac{Corr(EnsMean, Obs)}{Corr(EnsMean, Mem)}, \tag{2}$$

where $Corr(EnsMean, Mem)$ denotes the average correlation between the ensemble mean and individual ensemble members. Thus RPC > 1 can be understood as a situation where the ensemble mean correlates better with the real world than with random members, which in turn implies that the forecast underestimates the predictability of the real world. Equation 2 makes it clear that for a statistically perfect forecast (one where observations are indistinguishable from a random ensemble member), $RPC = 1$. Further discussion on RPC can be found in Scaife and Smith (2018) and Strommen and Palmer (2019).

2.5. The "Signal-Plus-Noise" Statistical Model

We use the idealized "signal-plus-noise" statistical model defined in Siegert et al. (2016), which we refer to as the SN-model for short. It assumes the forecast signals are linear and Gaussian. The reader should refer to their paper for extensive discussion. Here we simply recap the basic details we need.

Let y_t be the NAO index of the real world, and $x_{t,k}$, $k = 1, \ldots R$ be the NAO indices of an ensemble forecast of y with R members. If the NAO indices have been defined so as to have zero mean, the SN-model supposes that

$$y_t = s_t + \epsilon_t$$

$$x_{t,k} = \beta s_t + \eta_{t,k}.$$

19448007, 2023, 14, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL103710 by Oxford University, Wiley Online Library on [03/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Here s_t, ϵ_t and η_t are all independent, normally distributed variables with mean zero and standard deviations σ_s, σ_ϵ, σ_η, and β is a constant representing the sensitivity of the forecasts to the observations. One can interpret this as decomposing the observed NAO y_t into a predictable signal s_t, and an unpredictable noise term ϵ_t. The forecast attains skill by capturing a proportion βs_t of s_t, and has its own noise given by η_t. The ensemble members are assumed to be completely exchangeable with each other, exhibiting both the same signal and same level of noise.

It will be useful to note that the independence assumptions imply that $\mathrm{Var}(y) = \sigma_s^2 + \sigma_\epsilon^2$ and $\mathrm{Var}(x) = \beta^2\sigma_s^2 + \sigma_\eta^2$. Siegert et al. (2016) also derive a formula for the RPC of the SN-model in their Appendix B, which in the limit of infinitely many ensemble members becomes

$$RPC_{SN} = \frac{1}{\beta}\frac{\sqrt{\beta^2\sigma_s^2 + \sigma_\eta^2}}{\sqrt{\sigma_s^2 + \sigma_\epsilon^2}}. \tag{3}$$

Thus RPC > 1 occurs in this model as a result of either a small signal ($\beta < 1$) or excessive forecast variance (which when $\beta = 1$ happens if $\sigma_\eta > \sigma_\epsilon$).

To compare the forecast data to the SN-model behavior, we fit the free parameters of the SN-model to UKMO and ECMWF forecast data. To do so, we used the "moment estimator method" described in Appendix C of Siegert et al. (2016). The estimated values of (σ_s, σ_ϵ, σ_η, β) are (0.79, 0.61, 0.99, 0.27) for UKMO data, and (0.60, 0.80, 0.98, 0.37) for ECMWF data. These values will be discussed in Section 4.1. Because this discussion is not central to the paper, error-bars are omitted, but one can infer from the uncertainty estimates in Siegert et al. (2016) that the differences between UKMO and ECMWF parameters are likely not significant.

3. Results Using the Idealized Statistical Model

3.1. Numerical Analysis

In order to understand how reliability relates to RPC in the SN-model, we proceed as follows. We first fix the "observational" parameters σ_s and σ_ϵ to be the UKMO estimates from Section 2.5. We then pick random values of the "forecast" parameters β and σ_η uniformly between 0.10 and 10. These parameters are used to generate an observational timeseries y and 50 ensemble member timeseries x_k, each of length $N = 1{,}000$. For each such pair of observations and ensemble, we compute the slope of reliability diagrams corresponding to the upper and lower tercile events, along with the RPC value. The large sample size of 1,000 reduces the sampling variability and helps highlight general patterns. Figures 1a and 1b show how these metrics vary as a function of both β and $\sigma_\eta^2/\sigma_\epsilon^2$ in the case of the upper tercile event. Note that because the SN-model is linear, reliability diagrams for upper and lower terciles are always identical. Large (small) values of β are interpreted as the signal being large (small) in the forecast, while large (small) values of $\sigma_\eta^2/\sigma_\epsilon^2$ are interpreted as the forecast members exhibiting more (less) unpredictable noise than the real world. We refer to this latter ratio as the noise-ratio for short.

Several points can be inferred from Figure 1. First, it can be seen that the reliability of the slope varies monotonically with both β and the noise-ratio. Three essential parameter regimes can be identified, corresponding to the value of the slope S: $S < 1$ (overconfident), $S > 1$ (underconfident) and $S = 1$ (perfect reliability). An example reliability diagram from the $S < 1$ regime is shown in Figure 1c ($\beta = 2$, $\sigma_\eta = \sigma_\epsilon/2$), and an example for the $S > 1$ regime in Figure 1d ($\beta = 0.5$, $\sigma_\eta = 2\sigma_\epsilon$). In order for the forecast model to be perfectly statistically calibrated, it is necessary for both β and the noise-ratio to be 1. However, Figure 1a shows that perfect reliability can be obtained for a non-perfect forecast model through a compensation of errors: too much (little) noise can be balanced by an overly strong (weak) signal, and vice versa. This explains why the $S = 1$ regime sits on the diagonal.

Second, comparing Figures 1a and 1b strongly suggests that the parameter regimes defined by $RPC < 1$, $RPC > 1$ ("paradox") and $RPC = 1$ are *the same* as those defined using the reliability slope. In other words, the model parameters that lead to a reliability slope $S > 1$ are the same parameters that give $RPC > 1$. In fact, in the next section we will prove this statement under the assumptions of a sufficiently large sample and ensemble size. This means that in the SN-model, the "paradox" can be precisely understood as an instance of a forecast with reliability slope $S > 1$, thereby justifying the interpretation in Eade et al. (2014).

A final point of note concerns the impact of sampling variability. The artifacts in the contour of Figure 1a suggest that sampling variability remains non-trivial even with a sample size of $N = 1{,}000$. This can be seen in the two

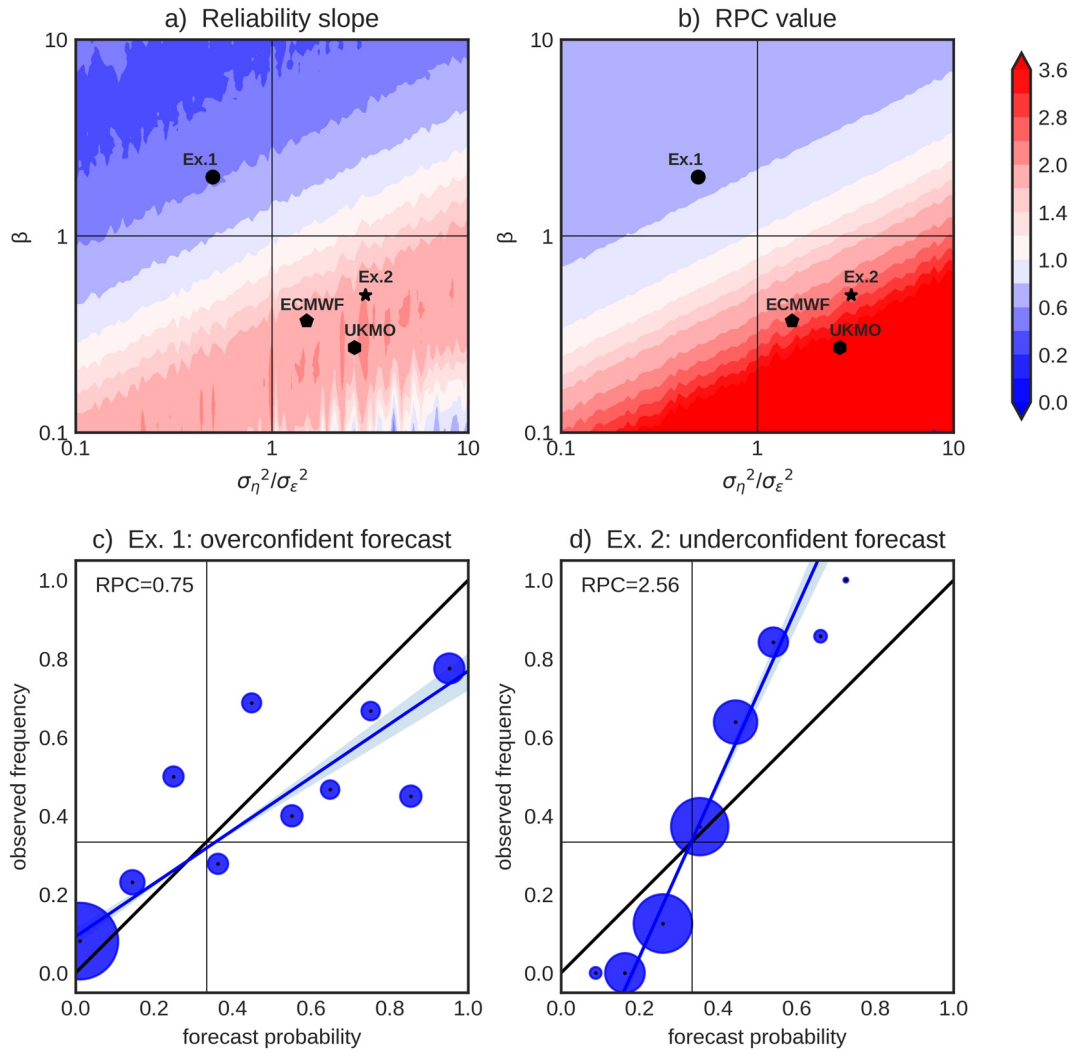

Figure 1. Reliability and ratio-of-predictable component (RPC) estimates using the SN-model. In (a) a contour plot of the slope of the reliability diagram across a range of SN-model parameters. In (b) the same but for RPC. In (c and d), example reliability diagrams obtained for a specific choice of parameters corresponding to an overconfident and underconfident forecast respectively. The sizes of the blue dots are proportional to the number of samples available in that bin; the blue line is the linear fit and the blue shading gives the 95% confidence interval of this linear fit. In (a and b), the location of the two examples (black dot and star) as well as the UK Met Office model and European Centre for Medium-range Weather Forecasts fits (black hexagon and pentagon) have been marked.

example diagrams (c) and (d), showing deviations from linearity that are necessarily due to sampling variability alone. The effect is especially big in the bottom right corner of Figure 1a, corresponding to the limiting case where ensemble members are purely noise-driven, implying that reliability cannot be sensibly assessed unless the forecast has sufficient skill. By comparison with Figure 1b, the sample size of 1000 appears sufficient to eliminate sampling variability for RPC estimates, even in the noise-driven limit case. However, in regions closer to the diagonal, the sampling variability of the reliability slope is small enough to easily assess which parameter-regime one is in.

3.2. Theoretical Analysis

We sketch a proof of the equivalence $S > 1 \Leftrightarrow RPC > 1$ in the SN-model, under the assumption that (a) the ensemble size is large enough that the ensemble mean $\hat{x}_t \approx \beta s_t$ (i.e., noise is completely eliminated), and (b) the sample size is large enough that sample estimates (of e.g., variances) equal the true underlying population values. For simplicity we also assume $\beta > 0$, that is, that the forecasts have non-zero skill. The sketch assumes the event

19448007, 2023, 14, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL103710 by Oxford University, Wiley Online Library on [03/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19448007, 2023, 14, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL103710 by Oxford University, Wiley Online Library on [03/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

definition E is the upper tercile: at the end we indicate why the same argument accounts for an arbitrary upper/lower percentile.

It is possible to derive exact formulas for the reliability curve (i.e., P_r as a function of P_f) which show that the curve is a strictly increasing "sigmoid" whose growth rate is determined by the *RPC*. These formulas exhibit degeneracies when the ensemble mean correlation is very close to 1, which in the SN-model happens when $\sigma_s \gg \sigma_\epsilon$. The sketch which follows is essentially correct away from this region and captures the key ideas. Complete details can be found in Supporting Information S1.

First note that given the assumption that ensemble mean correlations are not close to 1, the reliability curve will approximately pass through the point (1/3, 1/3). Intuitively, a forecast probability of 1/3 corresponds to a forecast which detects no predictable signals and which therefore gives us no knowledge about the value of y_t. Consequently, y_t is roughly speaking expected to be a random draw from its climatology, which lands in the upper tercile with probability 1/3, as desired.

Next, we will show that $RPC > 1$ if and only if the reliability curve passes through the point (1/2, L) for some $L > 1/2$, that is, a point above the diagonal. By definition, $L = \mathbb{P}(y_t \text{ satisfies } E | P_f = 0.5)$, where P_f is the forecast probability. Because y_t is normally distributed with variance $\sigma_s^2 + \sigma_\epsilon^2$, its upper tercile is defined by $y_t > \lambda\sqrt{\sigma_s^2 + \sigma_\epsilon^2}$, where $\lambda \approx 0.431$ defines the upper tercile threshold of the $\mathcal{N}(0, 1)$ distribution. Similarly, the upper tercile for the forecast distribution is defined by $x_{t,k} > \lambda\sqrt{\beta^2\sigma_s^2 + \sigma_\eta^2}$. Since ensemble members are normally distributed around the ensemble mean, $P_f = 0.5$ if and only if half the members exceed the upper tercile threshold, which happens if and only if the ensemble mean βs_t *equals* the forecast upper tercile threshold. Therefore,

$$L = \mathbb{P}\left(y_t = s_t + \epsilon_t > \lambda\sqrt{\sigma_s^2 + \sigma_\epsilon^2} | \beta s_t = \lambda\sqrt{\beta^2\sigma_s^2 + \sigma_\eta^2}\right). \tag{4}$$

This is the probability of ϵ_t exceeding a fixed threshold conditioned on the value of s_t. Since ϵ and s are independent, the conditional can be dropped, yielding

$$L = \mathbb{P}\left(\epsilon_t > \lambda\sqrt{\sigma_s^2 + \sigma_\epsilon^2} - \frac{\lambda}{\beta}\sqrt{\beta^2\sigma_s^2 + \sigma_\eta^2}\right). \tag{5}$$

Because ϵ_t is normally distributed with mean 0, this probability exceeds 0.5 if and only if the right-hand-side of the inequality in Equation 5 is less than 0. Rearranging and simplifying this implies

$$L > 0.5 \Leftrightarrow \frac{\sqrt{\beta^2\sigma_s^2 + \sigma_\eta^2}}{\beta\sqrt{\sigma_s^2 + \sigma_\epsilon^2}} > 1, \tag{6}$$

which by Equation 3 precisely says that RPC > 1.

We have shown that the reliability curve intersects the diagonal at $P_f = 1/3$ and is above the diagonal at $P_f = 1/2 \Leftrightarrow RPC > 1$. Since the reliability curve is strictly increasing, this already clarifies why $S > 1 \Leftrightarrow RPC > 1$. Similar arguments show that the curve is always below the diagonal for $P_f < 1/3$ and always above the diagonal for $P_f > 1/2$. Finally, one shows that the weights of the weighted regression are concentrated in the approximately linear part of the sigmoid with a peak for $1/3 < P_f < 1/2$, finishing the sketch.

The case where E is the lower tercile only requires a slight modification: the lower terciles are defined by the condition of being less than $-\lambda$ times the variance. The reversing of the inequality and the change of the sign cancel out in the end to give the same conclusion. If a different percentile C had been used to define E, then as long as $C < 0.5$ the exact same argument will work. If $C > 0.5$, the same argument can be applied to the lower percentile event $1 - C$ to establish the same claim by symmetry.

We note that an expression for the root mean-squared-error divided by the average ensemble spread can also be straightforwardly established under the same assumptions of large sample and ensemble size:

$$\frac{\text{RMSE}}{\text{Spread}} = \frac{\sqrt{(\beta - 1)^2\sigma_s^2 + \sigma_\epsilon^2}}{\sigma_\eta}, \tag{7}$$

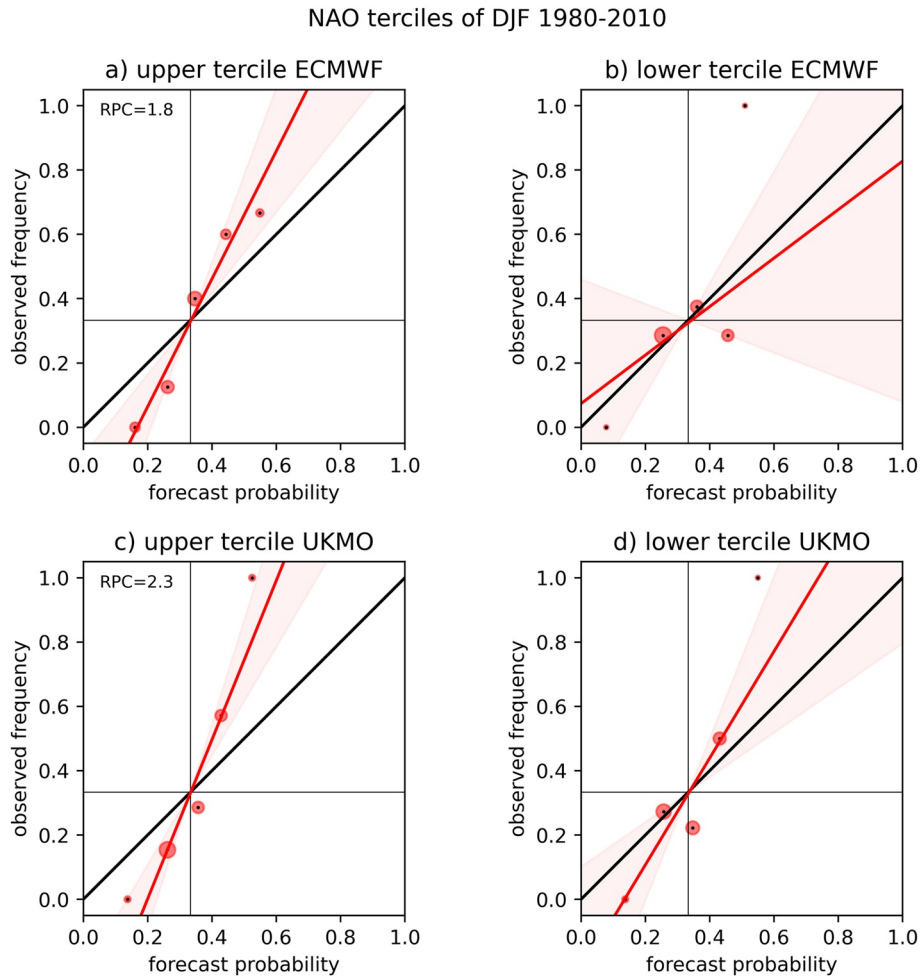

Figure 2. In (a and b), reliability diagrams of, respectively, upper and lower tercile DJF North Atlantic Oscillation forecasts by the European Centre for Medium-range Weather Forecasts ensemble, and in (c) and (d) the same but for UK Met Office model forecasts. The period covered is 1980–2010. The sizes of the red dots are proportional to the number of samples available in that bin; the thick red line is the linear fit and the red shading gives the 95% confidence interval of this linear fit. The "perfect reliability" diagonal (thick black line) is included for convenience.

where the precise meaning of the root-mean-square-error RMSE and "Spread" are as in Fortin et al. (2014) (see their equation (16)). Given a statistically perfect forecast this ratio equals 1 (Fortin et al., 2014), which can be easily verified in our case by setting $\beta = 1$ and $\sigma_\epsilon = \sigma_\eta$, the conditions required for perfect statistical calibration (Siegert et al., 2016). However, it is clear that for a non-perfect forecast, the relationship between this ratio and RPC (or the reliability slope) is not straightforward, suggesting that spread-error metrics are measuring forecast confidence in a fundamentally different manner to RPC and reliability diagrams.

4. Results Using Forecast Data

We consider in turn the modern period 1980–2010 and the full period 1900–2010. Note that a discussion of the challenges arising from sampling variability is reserved for Section 5.

4.1. The Period 1980-2010

Figure 2 shows the reliability diagrams of seasonal winter NAO forecasts using both the upper and lower tercile events, for ECMWF and UKMO forecast data, using the period 1980–2010 ($N = 31$) for which they overlap. Thick red lines show the raw estimate, with shading indicating uncertainty.

19448007, 2023, 14, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL103710 by Oxford University, Wiley Online Library on [03/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

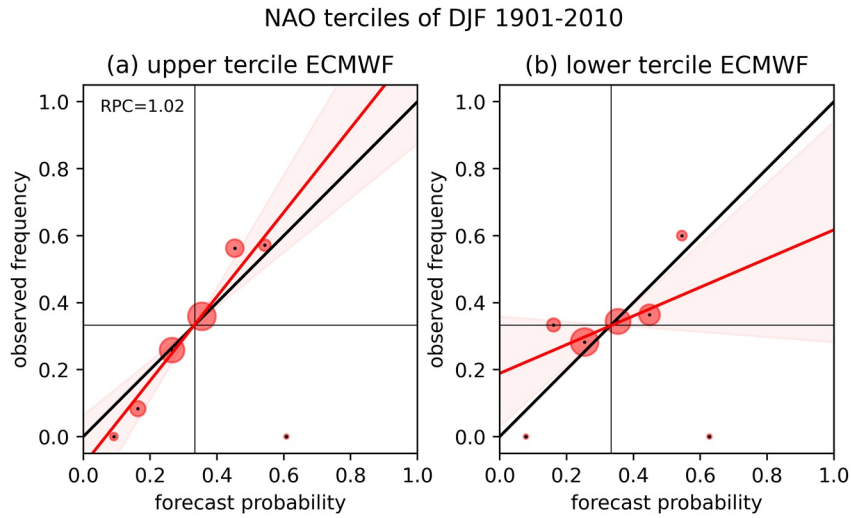

Figure 3. In (a and b), reliability diagrams of, respectively, upper and lower tercile DJF North Atlantic Oscillation forecasts by the European Centre for Medium-range Weather Forecasts ensemble. The period covered is 1901–2010. The sizes of the red dots are proportional to the number of samples available in that bin; the thick red line is the linear fit and the red shading gives the 95% confidence interval of this linear fit. The "perfect reliability" diagonal (thick black line) is included for convenience.

In diagrams (a), (c), and (d), the NAO reliability slope exceeds 1, and robustly so for the upper tercile event. In (b), the lower tercile event for ECMWF, the uncertainty is too great to assess the reliability, but the overall assessment of both NAO forecasts is that they are unreliable and underconfident. Given the conclusions of the previous section, this is consistent with the presence of the "signal-to-noise paradox" of these forecasts, with the UKMO and ECMWF exhibiting an RPC of 2.3 and 1.8 respectively. Therefore, despite the large uncertainties in the exact quantitative estimates of several of the reliability slopes in Figure 2, these diagrams give the same qualitative conclusion as analysis based on RPC.

The equivalence of reliability and RPC in the previous section assumed the linear and Gaussian SN-model. The only indication of a violation of linearity/Gaussianity here is in the discrepancy being between the upper and lower ECWMF terciles, though the uncertainty of the lower tercile slope is large enough to still be consistent with the SN-model.

In order to allow for a qualitative comparison between UKMO and ECMWF, their positions in $\left(\beta, \sigma_\eta^2/\sigma_\epsilon^2\right)$-space have been marked on Figures 1a and 1b, though we remind the reader that the uncertainty in the parameters means this comparison should be treated cautiously. The ECMWF forecasts appear to be slightly more reliable than UKMO, exhibiting both a higher β and a noise-ratio closer to 1. However, this is at the expense of less skill than UKMO, due to a smaller overall signal σ_s. The differing values of σ_s and σ_ϵ for the two datasets (Section 2.5) may seem puzzling, given the same observations are used for each. However, as emphasized in Siegert et al. (2016), the decomposition of observations into s and ϵ is a statistical construct which is in no way independent of the forecast product being used. For example, if UKMO simulates a teleconnection missing in ECMWF, a parameter fit using UKMO will assign the variability associated with this teleconnection to the signal, while ECMWF would assign it to the noise.

4.2. The Period 1901-2010

Figure 3 shows the reliability diagrams for the ECMWF model covering the full 110-year period 1901–2010 ($N = 109$). While the upper tercile uncertainty crosses the diagonal, the face-value reliability slope indicates underconfidence, consistent with the modern period 1980–2010. On the other hand, the lower tercile interestingly indicates robust *overconfidence*, giving an overall impression of good reliability when considering both terciles jointly.

These results are consistent with the analysis in Weisheimer et al. (2019), which showed that RPC $\approx$ 1 in these forecasts when computed over the period 1901–2010. Their analysis further showed pronounced decadal

19448007, 2023, 14, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL103710 by Oxford University, Wiley Online Library on [03/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19448007, 2023, 14, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL103710 by Oxford University, Wiley Online Library on [03/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

variability in both skill and RPC, with the modern period standing out as an era of relatively high skill and RPC values. These reliability diagrams, if taken at face value, complement Weisheimer et al. (2019) by indicating that the main source of both skill and high RPC values are forecasts of positive NAO events, with forecasts of negative events being qualitatively different. In particular, the ECMWF underconfidence of positive NAO events and overconfidence of negative NAO events appear to be relatively consistent features across both the full period 1901–2010 and the modern period 1980–2010.

5. Discussion and Conclusions

We have shown, given the assumption of linearity/Gaussianity, that the "signal-to-noise paradox" corresponds precisely to a situation where upper/lower percentile forecasts have a reliability diagram with a slope exceeding 1. More precisely, by utilizing the linear statistical model of Siegert et al. (2016), we showed that given a large sample size and sufficiently many ensemble members, the RPC metric exceeds 1 if and only if the reliability slope exceeds 1: the higher the RPC, the steeper the slope, and vice versa. This justifies the interpretation given in Eade et al. (2014) of the "paradox" as a case of an "underconfident forecast", with confidence measured probabilistically using reliability diagrams. On the other hand, the ratio of RMSE over ensemble spread is not straightforwardly related to RPC, meaning that this interpretation does not hold if confidence is measured using spread-error ratios.

We also showed, using ECMWF and UKMO seasonal hindcasts, that tercile forecasts of the winter NAO do exhibit reliability diagrams with slopes exceeding 1, corroborating the implicit point in Eade et al. (2014) that the "signal-to-noise paradox" of NAO forecasts can be detected using reliability diagrams. Consideration of the ECMWF hindcast, which covers the full 20th century, suggests that the main source of forecast underconfidence comes from positive NAO forecasts, with negative NAO forecasts being more overconfident on average. This is consistent with the results of Weisheimer et al. (2017) which found that the ECMWF forecasts of negative NAO events were somewhat less skillful. Negative NAO events are associated with increased occurrence of blocking episodes (Woollings et al., 2008) which are often hard to predict. Asymmetries between positive and negative NAO forecasts indicates a violation of linearity/Gaussianity, and may arise due to the effects of skew (Stephenson et al., 2004), flow-dependent predictability (Ferranti et al., 2015; Frame et al., 2013; Matsueda & Palmer, 2018), or non-linearities and state dependence (Önskog et al., 2018, 2020) such as that expected from regime dynamics (Strommen, 2020). Our results suggest that taking into account such asymmetries may help shed light on the "paradox."

The clear limitation to the use of reliability diagrams to assess seasonal mean forecasts, such as the winter NAO, is the uncertainty arising from sampling variability. This uncertainty has two sources. First, given the 1980–2010 hindcast sample size of 31, each forecast-probability bin was found to contain somewhere between 3 and 10 samples, which is clearly insufficient to robustly estimate the conditional observed frequency. Second, the "paradox" has the effect of clustering forecast probabilities close to 50%, meaning there are few cases of extreme forecast probabilities available, especially high-probability cases. The resulting uncertainty means that reliability diagrams based on a typical hindcast sample size of 30–40 years can only sensibly be used for qualitative, rather than quantitative, assessment. Reducing this uncertainty ultimately requires more samples, in order to add more points on the reliability diagram. Increased ensemble size beyond 40–50 would be expected to only slightly improve estimates of P_f, and hence not compensate for a small sample size (Leutbecher, 2019). Unfortunately, longer hindcasts spanning the 20th century currently only exist for the ECMWF model.

The fact that reliability diagrams are particularly sensitive to sampling variability is well known, and several "tactics" have become standard for overcoming this. For example, when assessing reliability of seasonal forecasts for a particular region (such as the UK), it is common to treat forecasts for each individual gridpoint in the region as independent instances of the regional forecast (Weisheimer & Palmer, 2014). While this has the effect of dramatically increasing the sample size, the assumptions will clearly often fail: neighboring gridpoints are not independent and the presence of orography means individual gridpoints may not be representative of the region as a whole. The large uncertainties in the 1980–2010 winter NAO reliability estimates (Figure 2) may seem less unfavorable in this light. We also note that uncertainties in RPC estimation are typically considerable: in cases of low forecast skill (ensemble mean correlations <0.4) these uncertainties can easily be large enough to make it impossible to assess if the RPC is greater or less than 1 (Strommen & Palmer, 2019).

19448007, 2023, 14, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL103710 by Oxford University, Wiley Online Library on [03/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Nevertheless, it is natural to ask if tactics similar to the use of gridpoint forecasts can be used to more robustly assess the reliability of winter NAO forecasts. One possibility is to use forecasts of the December, January and February NAO separately. This was explored, but found to pose challenges, since forecast skill was found to not be uniform across each month, with December showing little to no skill. It is therefore not immediately clear how to relate reliability of the pooled monthly forecasts to reliability of the seasonal mean forecast. Other future avenues of exploration might include pooling forecasts of multiple principal components beyond just the first.

In conclusion, despite the limitations imposed by sampling variability, we propose that reliability diagrams of seasonal means can provide a useful complementary view of the "signal-to-noise paradox," and more broadly contribute to the qualitative assessment of seasonal forecasts. In particular, exploration of both upper and lower percentile forecasts seems valuable as an easy way to help identify the largest contributors to the "paradox" in a way that the raw RPC cannot. It would be interesting to assess if the relationship between reliability diagrams and RPC holds in more non-linear models, such as the regime-based one of Strommen and Palmer (2019).

Data Availability Statement

The NAO hindcast timeseries and python code used to create the figures of this paper are freely available via Zenodo: https://zenodo.org/badge/latestdoi/611695189. ASF20C data is freely available on CEDA (Weisheimer & O'Reilly, 2020). ERA20C data is available via the ECMWF Archive Catalog: https://apps.ecmwf.int/archive-catalogue/ The catalog is public but to download the data you need to request access from ECMWF.

Acknowledgments
KS gratefully acknowledges funding from the Thomas Philips and Jocelyn Keene Junior Research Fellowship, Jesus College. MM acknowledges funding from the Met Office Academic Partnership (MOAP), which funded their work across the summers of 2021 and 2022. HC acknowledges NERC grant NE/P018238/1.

References

Athanasiadis, P. J., Bellucci, A., Scaife, A., Hermanson, L., Materia, S., Sanna, A., et al. (2017). A multisystem view of wintertime NAO seasonal predictions. *Journal of Climate*, *30*(4), 1461–1475. https://doi.org/10.1175/jcli-d-16-0153.1

Bröcker, J., Charlton-Perez, A. J., & Weisheimer, A. (2023). A statistical perspective on the signal-to-noise paradox. *Quarterly Journal of the Royal Meteorological Society*, *149*(752), 911–923. https://doi.org/10.1002/qj.4440

Bröcker, J., & Smith, L. A. (2007). Increasing the reliability of reliability diagrams. *Weather and Forecasting*, *22*(3), 651–661. https://doi.org/10.1175/waf993.1

Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Eade, R., Robinson, N., et al. (2016). Skilful predictions of the winter North Atlantic Oscillation one year ahead. *Nature Geoscience*, *9*(11), 809–814. https://doi.org/10.1038/ngeo2824

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., & Robinson, N. (2014). Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, *41*(15), 5620–5628. https://doi.org/10.1002/2014GL061146

Ferranti, L., Corti, S., & Janousek, M. (2015). Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, *141*(688), 916–924. https://doi.org/10.1002/qj.2411

Fortin, V., Abaza, M., Anctil, F., & Turcotte, R. (2014). Why should ensemble spread match the RMSE of the ensemble mean? *Journal of Hydrometeorology*, *15*(4), 1708–1713. https://doi.org/10.1175/jhm-d-14-0008.1

Frame, T., Methven, J., Gray, S., & Ambaum, M. (2013). Flow-dependent predictability of the North Atlantic jet. *Geophysical Research Letters*, *40*(10), 2411–2416. https://doi.org/10.1002/grl.50454

Johnson, C., & Bowler, N. (2009). On the reliability and calibration of ensemble forecasts. *Monthly Weather Review*, *137*(5), 1717–1720. https://doi.org/10.1175/2009mwr2715.1

Leutbecher, M. (2019). Ensemble size: How suboptimal is less than infinity? *Quarterly Journal of the Royal Meteorological Society*, *145*(S1), 107–128. https://doi.org/10.1002/qj.3387

Matsueda, M., & Palmer, T. (2018). Estimates of flow-dependent predictability of wintertime Euro-Atlantic weather regimes in medium-range forecasts. *Quarterly Journal of the Royal Meteorological Society*, *144*(713), 1012–1027. https://doi.org/10.1002/qj.3265

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., et al. (2011). *The new ECMWF seasonal forecast system (System 4)* (Vol. 49). European Centre for Medium-Range Weather Forecasts.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, *12*(4), 595–600. https://doi.org/10.1175/1520-0450(1973)012<0595:anvpot>2.0.co;2

Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *26*(1), 41–47. https://doi.org/10.2307/2346866

Önskog, T., Franzke, C. L., & Hannachi, A. (2018). Predictability and non-Gaussian characteristics of the North Atlantic Oscillation. *Journal of Climate*, *31*(2), 537–554. https://doi.org/10.1175/jcli-d-17-0101.1

Önskog, T., Franzke, C. L. E., & Hannachi, A. (2020). Nonlinear time series models for the North Atlantic Oscillation. *Advances in Statistical Climatology, Meteorology and Oceanography*, *6*(2), 141–157. https://doi.org/10.5194/ascmo-6-141-2020

Poli, P., Hersbach, H., Dee, D. P., Berrisford, P., Simmons, A. J., Vitart, F., et al. (2016). ERA-20C: An atmospheric reanalysis of the twentieth century. *Journal of Climate*, *29*(11), 4083–4097. https://doi.org/10.1175/JCLI-D-15-0556.1

Scaife, A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R., Dunstone, N., et al. (2014). Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, *41*(7), 2514–2519. https://doi.org/10.1002/2014gl059637

Scaife, A., & 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

Siegert, S., Stephenson, D. B., Sansom, P. G., Scaife, A., Eade, R., & 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. https://doi.org/10.1175/jcli-d-15-0196.1

Stephenson, D., Hannachi, A., & O'Neill, A. (2004). On the existence of multiple climate regimes. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *130*(597), 583–605. https://doi.org/10.1256/qj.02.146

Strommen, K. (2020). Jet latitude regimes and the predictability of the North Atlantic Oscillation. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 2368–2391. https://doi.org/10.1002/qj.3796

Strommen, K., & 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

Weisheimer, A., Decremer, D., MacLeod, D., O'Reilly, C., Stockdale, T. N., Johnson, S., & 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., & O'Reilly, C. (2020). Initialised seasonal forecast of the 20th century. [Dataset]. Centre for Environmental Data Analysis. Retrieved from https://catalogue.ceda.ac.uk/uuid/6e1c3df49f644a0f812818080bed5e45

Weisheimer, A., & Palmer, T. (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., Schaller, N., O'Reilly, C., MacLeod, D. A., & Palmer, T. (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

Woollings, T. J., Hoskins, B., Blackburn, M., & Berrisford, P. (2008). feb). A new Rossby wave-breaking interpretation of the North Atlantic Oscillation. *Journal of the Atmospheric Sciences*, *65*(2), 609–626. https://doi.org/10.1175/2007JAS2347.1

19448007, 2023, 14, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL103710 by Oxford University, Wiley Online Library on [03/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License