When and where do ECMWF seasonal forecast systems exhibit anomalously low signal-to-noise ratio?

Article

Accepted Version

When and where do ECMWF seasonal forecast systems exhibit anomalously low signal-to-noise ratio?

Andrew J. Charlton-Perez,[a]*, Jochen Bröcker [a,b], Timothy N. Stockdale[c], Stephanie Johnson [c]

[a]*Dept. of Meteorology, Univ. of Reading, Reading, UK*

[b]*Dept. of Mathematics and Statistics, Univ. of Reading, Reading, UK*

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

*Correspondence to: Department of Meteorology, Univ. of Reading, Whiteknights, Reading, Berks, RG6 6BB, a.j.charlton-perez@reading.ac.uk

Seasonal predictions of wintertime climate in the Northern Hemisphere mid-latitudes, while showing clear correlation skill, suffer from anomalously low signal-to-noise ratio. The low signal-to-noise ratio means that forecasts need to be made with large ensemble sizes and require significant post-processing to correct the forecast distribution. In this study, a recently introduced statistical model of seasonal climate predictability is adapted so that it can be used to examine the signal-to-noise ratio in two versions of the ECMWF seasonal forecast system. Three novel features of the low signal-to-noise ratio are revealed. The low signal-to-noise ratio is present only for forecasts initialised on November 1st and not for forecasts initialised on December 1st. The low signal-to-noise ratio is predominantly a feature of the lower and middle troposphere and is not present in the stratosphere. The low signal-to-noise ratio is linked to low signal amplitude of the forecast systems in early winter. Future studies attempting to examine the signal-to-noise ratio should focus on the extent to which this early winter variability is predictable.

Key Words: SIGNAL-TO-NOISE, SEASONAL PREDICTION, NORTHERN ANNULAR MODE, ARCTIC OSCILLATION

Received . . .

1. Introduction and Motivation

Recent studies have clearly shown that skillful prediction of climate in the North Atlantic is possible on season ahead (Scaife *et al.* 2014; Stockdale *et al.* 2015) or even year ahead timescales (Dunstone *et al.* 2016). The increase in the capability of seasonal forecasting

systems to skilfully predict North Atlantic climate has been partially attributed to the improved representation of stratospheric processes in seasonal forecasting systems (Scaife *et al.* 2016). The polar stratosphere can provide a pathway for remote sources of predictability, such as ENSO or the QBO, to the North Atlantic (e.g. Kidston *et al.* 2015). Comparison of high and low-top models shows that high-top models are better able to capture the lower stratospheric response to ENSO and QBO variability (Butler *et al.* 2016). The importance of stratospheric variability in producing skillful forecasts of North Atlantic climate is quantified by Domeisen *et al.* (2015) and Scaife *et al.* (2016), who show that the NAO forecast skill vanishes when ensemble members that produce Sudden Stratospheric Warming (SSW) events are excluded from the forecast ensemble.

An interesting aspect of the skill of NAO or Arctic Oscillation (AO) forecasts in both the Met Office and European Centre for Medium-Range Weather Forecasts (ECMWF) systems is the apparent signal-to-noise paradox in these forecasts (Eade *et al.* 2014). This paradox is clearly illustrated in Fig. 1 of Scaife *et al.* (2014) and Fig. 1 of Stockdale *et al.* (2015), the ensemble mean forecast of the seasonal climate state is highly correlated with the observed state but has unexpectedly low variance. Further evidence for the signal-to-noise paradox, and the extent to which this is a ubiquitous feature of climate models, is reviewed extensively by Scaife and Smith (2018).

For seasonal predictions, the unexpectedly low signal-to-noise ratio can be quantified either by the Ratio of Predictable (RPC) components diagnostic Eade *et al.* (2014) or by explicitly quantifying the signal-to-noise ratio in the forecast system and observations (Siegert *et al.* 2016). The Bayesian framework introduced by Siegert *et al.* (2016) makes it possible to state with high confidence (posterior probability of 0.99) that the Met Office GloSea5 system 'underestimates the predictability of the real-world'. Baker *et al.* (2018) use the RPC diagnostic to examine forecasts from seven different seasonal prediction systems and show low signal-to-noise ratio can be clearly diagnosed in four systems, indicating that this is a widespread and common problem for seasonal prediction. For ECMWF System 4, Baker *et al.* (2018) calculate that the RPC diagnostic does not indicate anomalously low signal-to-noise ratio for the NAO index. However, Weisheimer *et al.* (2018) show that, at least over Greenland and calculated over a longer period, RPC is significantly greater than one in both System 4 and System 5. Weisheimer *et al.* (2018) also show the large sampling variability of the RPC diagnostic when applied to short hindcast sets. For the Arctic Oscillation, Stockdale *et al.* (2015) show that ECMWF System 4 has anomalously low signal-to-noise ratio for DJF forecasts.

While both Eade *et al.* (2014) and Siegert *et al.* (2016) provide post-processing techniques that can be used to re-calibrate forecasts to account for the low signal-to-noise ratio of the forecast systems for practical use, understanding the origin of the low signal-to-noise ratio would help to determine if and how it can be eliminated. In this study, we adapt the methodology of Siegert *et al.* (2016) in order to study the signal-to-noise ratio of two ECMWF seasonal forecasting systems in more detail. Our key objective is to understand over which time periods and in which part or parts of the atmosphere the low signal-to-noise ratio emerges.

2. Methodology

2.1. Statistical Model

To describe the statistical model, we first begin by defining observations of the seasonal mean climate as $\{Y(t), t = 1, 2, \ldots\}$ and the corresponding forecasts as $\{\mathbf{X_k}(t), t = 1, 2, \ldots\}$. In particular, the ensemble members at time t will be denoted by $X_1(t), \ldots, X_K(t)$ and there are K members in each ensemble. The starting point for our analysis is splitting of the ensemble members and the observations into what might be called the *predictable part* and the *noise*. $\mathcal{I}(t)$ is the information available to the forecaster at the point when the forecast $\mathbf{X}(t)$ is issued (roughly speaking this will be t minus the lead time). We then have the decomposition

$$X_k(t) = U(t) + \eta_k(t)$$

where we define $U(t) := \mathbb{E}(X_k(t)|\mathcal{I}(t))$ and $\eta_k(t) := X_k(t) - \mathbb{E}(X_k(t)|\mathcal{I}(t))$. Our first assumption is that $U(t)$ is the same for all ensemble members. Note that $\eta_k(t)$ and $U(t)$ will be uncorrelated. Likewise, we decompose

$$Y(t) = V(t) + \epsilon(t)$$

with $V(t) := \mathbb{E}(Y(t)|U(t))$, and we can interpret $V(t)$ as the predictable part of the observation and $\epsilon(t)$ as the noise. Note that the predictable part $V(t)$ is a function of $U(t)$, and that $\epsilon(t)$ is uncorrelated with the predictable part and in fact also with $U(t)$ itself. Our second assumption is that $\epsilon(t)$ is also uncorrelated with $\eta_k(t)$ for all $k = 1, \ldots, K$.

We define the signal-to-noise ratios of the observations and the forecast as

$$\mathrm{SNR}_y = \sqrt{\frac{\mathbb{E}(V(t)^2) - \mathbb{E}(V(t))^2}{\mathbb{E}(\epsilon(t)^2)}},$$
$$\mathrm{SNR}_x = \sqrt{\frac{\mathbb{E}(U(t)^2) - \mathbb{E}(U(t))^2}{\mathbb{E}(\eta_k(t)^2)}},$$

respectively. Our third assumption is that these numbers do not depend on k or t. Our final assumption is that forecasts and observations share a common source of predictability. Mathematically, this means that $U(t)$ indeed carries some information about the verification $Y(t)$ and hence that $V(t) = \mathbb{E}(Y(t)|U(t))$ is not constant but exhibits some variability.

With these preparations, it is now a simple matter to show that the correlation coefficient R between the observations $Y(t)$ and the ensemble mean $m(t) = \sum_{k=1}^{K} X_k(t)$ satisfies the relation:

$$\begin{aligned} R(Y(t), m(t)) &= R(V(t), U(t)) \cdot \sqrt{\frac{K \cdot \mathrm{SNR}_x^2}{1 + K \cdot \mathrm{SNR}_x^2}} \cdot \sqrt{\frac{\mathrm{SNR}_y^2}{1 + \mathrm{SNR}_y^2}} \\ &\leq \sqrt{\frac{K \cdot \mathrm{SNR}_x^2}{1 + K \cdot \mathrm{SNR}_x^2}} \cdot \sqrt{\frac{\mathrm{SNR}_y^2}{1 + \mathrm{SNR}_y^2}} \end{aligned} \tag{1}$$

The signal to noise paradox arises if we assume that $\mathrm{SNR}_y \cong \mathrm{SNR}_x$, while the former is in fact slightly larger. As a consequence, the correlation between observations and the ensemble mean might turn out to be much larger than our incorrect assumption, in combination with the estimate in Equation (1), would suggest. To analyse the signal to noise paradox statistically, Siegert *et al.* (2016) use a more specific model for seasonal forecast data. In essence, they assume that $U(t), \epsilon(t), \eta_1, \ldots, \eta_K(t)$ are normal random variables that have the correlation properties derived above but in addition are independent over time. Further, $V(t)$ is essentially a linear function of $U(t)$. These are highly idealistic assumptions, but pending the development of a more powerful statistical methodology, we will use a variant of this model in the present study.

Siegert *et al.* (2016) fit this model to seasonal forecast data using Markov chain Monte Carlo (MCMC) method. While this method has many advantages and is widely used, it is necessary to specify a prior distribution for the parameters since the method is an approximation of a fully Bayesian analysis. Since a key aim of this study is to fit the model over a range of different vertical levels and seasons, careful specification of a prior for the large numbers of cases considered becomes difficult. Instead, a new method is developed that allows us to fit the same statistical model to sets of forecasts from two ECMWF seasonal forecasting systems. Essentially, we will use a Maximum Likelihood approach. This can also be interpreted as finding the most probable parameter values in a Bayesian setting, albeit with uninformative or "flat" priors for a certain parametrisation of the problem, namely through the covariance matrix M and parameter σ_m^2 defined below. It needs to be stressed that if a problem is reparametrised, the priors for the former parameters imply priors for the new parameters that are not flat. The equivalence between Maximum Likelihood estimates and most probable parameters

for a given parametrisation is only valid whenever a flat prior is chosen for that parametrisation. Siegert *et al.* (2016) found that specifying any sort of informative prior had a strong (but, in our view, unwarranted) influence on the parameters, given the small hindcast set typically available from most seasonal forecasting systems. We also found that the regularising effect that comes with informative priors was not necessary in our case to stabilise the estimates. One further advantage of this method, is that it is slightly computationally less intensive than the MCMC analysis, even with bootstrapping of model parameters as discussed below. We first rewrite the statistical model as:

$$Y(t) = \mu_y + \beta_y S(t) + \epsilon O(t) \tag{2}$$

$$X_k(t) = \mu_x + \beta_x S(t) + \eta P_k(t) \qquad \text{for } k = 1, \ldots, K \tag{3}$$

where $S(t), O(t), P_1(t), \ldots, P_K(t)$ are independent standard normal random variables that are also independent over time (i.e. for different t). The parameters of the model are $\mu_y, \mu_x, \beta_y, \beta_x, \epsilon$ and η which represent the observational mean, model mean, observational signal amplitude, forecast signal amplitude, observational noise amplitude and forecast noise amplitude respectively. We note here that the signs of ϵ and η can be changed separately or jointly; we will therefore consider only nonnegative estimates of these two parameters. Secondly, β_y and β_x can be replaced simultaneously with $-\beta_y$ and $-\beta_x$; this however is irrelevant as we are interested in quantities involving the magnitude of these parameters, only. We note that $\beta_1 \cdot \beta_2$, the correlation between Y and $\mathbf{X}$ still has a well defined sign.

It turns out to be convenient to transform the forecasts, $\mathbf{X}$ using *contrasts*. A K–dimensional contrast $\mathbf{w}$ is an element of $\mathbb{R}^K$ such that $\sum_{k=1}^K w_k = 0$. We let $\{\mathbf{w}^{(l)} \in \mathbb{R}^K, l = 1, \ldots K-1\}$ be a set of $K-1$ orthogonal and normalised contrasts (superscripts are used to distinguish between different contrasts). Such a set can contain at most $K-1$ elements as the space of K–dimensional contrast has dimension $K-1$. In section A we describe the algorithm used to find such a set of contrasts.

By means of the contrasts, we introduce new random variables $(m(t), \zeta_1(t), \ldots, \zeta_{K-1}(t))$ that are in one–to–one relationship with $(X_1(t), \ldots, X_K(t))$ by setting:

$$m(t) = \frac{1}{K} \sum_{k=1}^{K} X_k(t) \tag{4}$$

$$\zeta_l(t) = \langle \mathbf{w}^{(l)}, \mathbf{X}(t) \rangle \qquad \text{for } l = 1, \ldots, K-1 \tag{5}$$

where $\langle ., .. \rangle$ denotes the standard scalar product on $\mathbb{R}^K$. Clearly, $m(t)$ is the ensemble mean. The reason for transforming the variables using contrasts and introducing the new variables $Y(t), m(t), \zeta_1(t), \ldots, \zeta_{K-1}(t)$ is the rather simple dependence structure among them. Indeed, a simple calculation reveals that:

1. The variables $\zeta_l(t), l = 1, \ldots, K-1, t = 1, 2, \ldots$ are independent (both for different l and in time) and normally distributed with mean zero and variance η^2.
2. The two–dimensional variables $\mathbf{Z}(t) := (Y(t), m(t))^{\mathrm{T}}, t = 1, 2, \ldots$ are independent in time and from the $\zeta_l(t)$; further, they are normally distributed with mean $(\mu_y, \mu_x)^{\mathrm{T}}$ and some covariance matrix M.
3. The covariance matrix M has the form

$$M = \begin{pmatrix} \beta_y^2 + \epsilon^2 & \beta_y \beta_x \\ \beta_y \beta_x & \beta_x^2 + \sigma_m^2 \end{pmatrix} \tag{6}$$

where $\sigma_m^2 := \eta^2 / K$.

The statistical model in Equations(2,3) can be equivalently parametrised in terms of σ_m and M. We find maximum likelihood estimators of the parameters $\beta_y, \beta_x, \epsilon, \eta$ by first finding maximum likelihood estimates of M, σ_m and subsequently inverting Equation 6 and the relation $\sigma_m = \eta^2/K$. This gives indeed the maximum likelihood estimators for these parameters, except for certain combinations of M, σ_m which are not admissible; see the discussion in the next section. More details about this approach can be found in the supplementary material.

$$\hat{\sigma}_m^2 = \frac{1}{K(K-1)N} \sum_{l \leq K-1, t \leq N} \zeta_l^2(t) \tag{7}$$

$$\hat{M} = \frac{1}{N} \sum_{t \leq N} Z(t) Z^{\mathrm{T}}(t) - \left(\frac{1}{N} \sum_{t \leq N} Z(t) \right) \left(\frac{1}{N} \sum_{t \leq N} Z(t) \right)^{\mathrm{T}} \tag{8}$$

To analyse the signal-to-noise properties of each forecast system we wish to estimate the signal-to-noise ratio of the observations and the forecasts, defined as:

$$\mathrm{SNR}_y = \frac{\beta_y}{\epsilon},$$
$$\mathrm{SNR}_x = \frac{\beta_x}{\eta},$$

Estimators for these terms are:

$$\hat{\mathrm{SNR}}_y = \frac{\hat{M}_{1,2}}{\sqrt{\det(\hat{M}) - \hat{\sigma}_m^2 \hat{M}_{1,1}}} \tag{9}$$

$$\hat{\mathrm{SNR}}_x = \frac{1}{\sqrt{K}} \sqrt{\frac{\hat{M}_{2,2}}{\hat{\sigma}_m^2} - 1}, \tag{10}$$

Estimators for all other parameters of the model can be similarly derived from $\hat{M}$ and $\hat{\sigma}_m$.

2.2. *Bootstrapping*

Uncertainty estimates for the parameters of the model are derived by parametric bootstrapping of the estimators $\hat{M}$ and $\hat{\sigma}_m$. In essence, parametric bootstrapping means to draw samples from a distribution that is known up to parameters which are replaced by estimators, see for instance Schervish (1995). In the present example, the distributions of $\hat{M}$ and $\hat{\sigma}_m$ are known structurally except for the parameters M and σ_m^2. Namely, $\hat{M}$ follows a Wishart distribution with N degrees of freedom and scale matrix M which we estimate by $\hat{M}$, so samples of $\hat{M}$ are drawn from a Wishart distribution with N degrees of freedom and scale matrix $\hat{M}$. Similarly, samples for $\hat{\sigma}_m^2$ are calculated by randomly drawing from a χ–square distribution with $(K-1) \cdot N$ degrees of freedom and multiplying with $\hat{\sigma}_m^2$. Note also that the resulting bootstrap samples for $\hat{M}$ and $\hat{\sigma}_m^2$ do not only have the correct distribution but are furthermore independent. This is the correct dependence structure in view of the fact that $\hat{M}$ and $\hat{\sigma}_m^2$ are independent, since they are functions of $\zeta(t)$ and $Z(t)$, respectively, which are independent. One reason for transforming the data using contrasts was to obtain this relatively simple dependence structure. It turns out that for some bootstrap resamples the SNR's are not well defined. Indeed, if either $\det(\hat{M}) \leq \hat{\sigma}_m^2 \hat{M}_{1,1}$ or $\hat{M}_{2,2} \leq \hat{\sigma}_m^2$, then $\hat{\mathrm{SNR}}_y$ or $\hat{\mathrm{SNR}}_x$, respectively, are not well defined. These cases are identified and discarded. The fraction of such cases varies between experiments but is never more than 0.1 % of samples and typically less. The reason that such cases occur is that there are instances of a covariance matrix M paired with a value for the parameter σ_m that are incompatible with the representation of M in Equation (6). If for instance $M_{22} < \sigma_m^2$, we get $\beta_x^2 < 0$ which is impossible; likewise $\det(M) < \sigma_m^2 M_{1,1}$ would imply $\epsilon^2 < 0$.

These are precisely the (M, σ_m^2)–pairs for which SNR_x or SNR_y, respectively, are not defined, and we will refer to such pairs as *not admissible*. We stress, however, that there is nothing unusual about pairs that are not admissible except that they cannot occur in connection with our model (2,3). But they *can*, with a small but nonvanishing probability, occur as values for the estimators (7, 8), even if the samples are drawn from our model (2,3). This is the reason why a small proportion of the parametric bootstrap resamples $(\hat{M}, \hat{\sigma}_m^2)$ cannot be translated into SNR resamples, as they are not admissible. Our approach of discarding these samples is equivalent to drawing $\hat{M}$ and σ_m^2 from Wishart resp. χ^2–distributions, *censored* to values that are admissible and thus also provide well defined values for the SNR's. A Bayesian approach as for instance used in Siegert *et al.* (2016) will not produce inadmissible resamples for the parameters if only because such resamples are deemed impossible apriori.

As an aside, if the estimates $\hat{M}$ and $\hat{\sigma}_m^2$ from Equations (7, 8) are inadmissible then they are in fact *not* the Maximum–Likelihood estimates for M and σ_m^2. Rather, the correct Maximum–Likelihood estimates will lie on the boundary of the set of admissible pairs, as a more refined analysis will reveal. This is not relevant for our purposes though, since this boundary is precisely the set of all (M, σ_m^2) for which ϵ^2 or η^2 (or both) are zero, and the SNR's are still not defined in this situation. For all calculations in the paper, sampling intervals for all parameters are calculated from 10^5 valid bootstrapped samples.

2.3. *Seasonal hindcasts and diagnostics*

We present results from the two most recent operational versions of the ECMWF seasonal forecasting system. SYS4 was based on IFS Cy36r4, run at TL255 spectral resolution with a linear reduced gaussian grid of 80km and 91 vertical levels, with a model top at 0.01 hPa. SYS4 was coupled to a 1x1 deg configuration of the NEMO ocean model, and sea-ice in the forecasts was specified based on sampling the previous 5 years, thus capturing the trend and uncertainty in sea-ice cover, but not predicting it. Atmospheric initial conditions for the re-forecasts are taken from ERA-interim (Dee *et al.* 2011) up to 2010 and the operational ECMWF analysis thereafter. The system is described in Molteni *et al.* (2011), and its performance in predicting the AO is described in Stockdale *et al.* (2015). SYS4 was operational for six years until October 2017, and was replaced by SEAS5 in November 2017. SEAS5 is based on IFS Cy43r1, run at Tco199 spectral resolution with a cubic octahedral reduced gaussian grid of 36 km and 91 vertical levels, with a model top at 0.01 hPa. SEAS5 is coupled to an ORCA025 configuration of the NEMO ocean and LIM2 interactive sea ice models. Atmospheric initial conditions for the re-forecasts are taken from ERA-interim up to 2016 and the operational ECMWF analysis thereafter, while ocean initial conditions are taken from OCEAN5 (Zuo *et al.* 2019). Details of SEAS5's configuration and performance, including stratospheric biases, are documented in Johnson *et al.* (2018).

We examine hindcasts for the November initialisations of SYS4 and November and December initialisations of SEAS5 for the period 1985 to 2016. The November initialisations have 51 ensemble members and the December initialisations have 25 ensemble members. The ERA-Interim dataset (Dee *et al.* 2011) is used as an estimate of the true, observed climate state. For both the hindcasts and observations, daily instantaneous values at 00UTC are used to derive averaged quantities.

Following Stockdale *et al.* (2015), two diagnostics are used to characterise large-scale climate variability in the Northern Hemisphere. At the surface, we calculate the Arctic Oscillation index (AO) from mean sea-level pressure (MSLP) data and at a range of pressure levels from 1000hPa to 10hPa we calculate the Northern Annular Mode (NAM) index from zonal mean geopotential height data. The AO pattern is defined as the first EOF of the monthly mean MSLP anomalies calculated between 1980 and 2010 for the ERA-Interim dataset. The daily AO index for the re-analysis and for each ensemble member is calculated by projecting daily MSLP anomalies onto the AO pattern. A similar procedure is used to calculate the NAM index, using zonal-mean geopotential height anomalies, following Baldwin and Thompson (2009). NAM patterns and projections are calculated separately for each pressure level.

The AO index calculated for the winter season (DJF) is shown in Fig. 1 for three different sets of forecasts from ECMWF hindcasts. In the figure, the ensemble mean is scaled by a constant factor so that variations in the ensemble mean can easily be observed. A

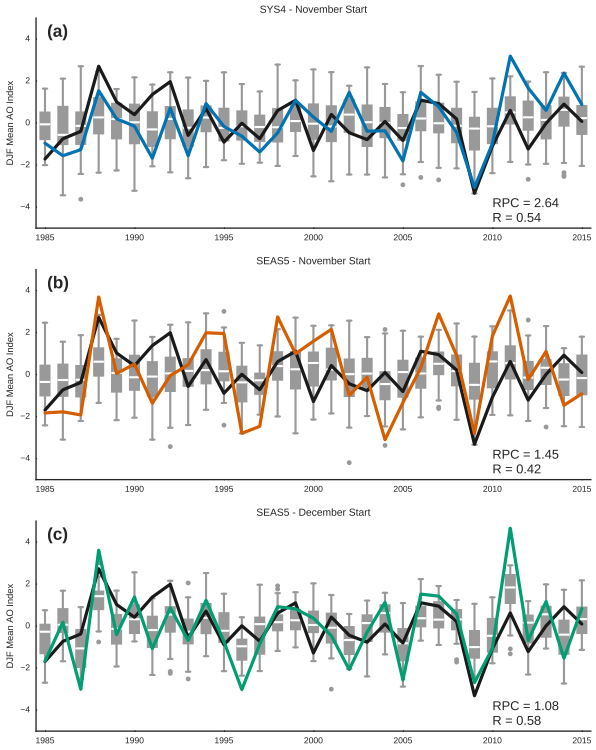

Figure 1. Observed (black line) and ensemble mean hindcast (coloured, solid line) for AO index for DJF from (a) forecasts initialised on November 1st using SYS4, (b) forecasts initialised on November 1st using SEAS5 and (c) forecasts initialised on December 1st using SEAS5. In (a) and (b), the ensemble mean is scaled by a factor of 6 and in (c) by a factor of 3 for ease of viewing the ensemble mean variation. The correlation coefficient between the ensemble mean and the observations and the Ratio of Predictable Components is reported at the bottom right of each plot. The distribution of forecasts for each year is shown by the grey box plots, with the median plotted in the white line, the inter-quartile range (IQR) shown in the box and the whiskers show the range of the data. Outliers are plotted as grey dots if they are more than 1.5 times the IQR away from the upper or lower edge of the box.

constant factor of 6 is used in panels (a) and (b) and 3 in panel (c). For forecasts initialised on 1st November, for both SYS4 and SEAS5, there is a moderately high correlation between the observed AO index and ensemble mean hindcast for this period (0.54 for SYS4 and 0.42 for SEAS5) but the size of variations in the ensemble mean is very small. The boxplots showing each ensemble member illustrate the large ensemble spread. As a first estimate of signal-to-noise biases in each hindcast set, the Ratio of Predictable Components (Eade *et al.* 2014) is shown for each combination of hindcasts and observations in the bottom right of each panel. For both the SYS4 and SEAS5 hindcasts, the RPC is larger than 1, suggesting over dispersion in these two hindcast sets. The remainder of the paper will use the statistical method outlined above to explore the origin of this anomalously low signal-to-noise ratio in the ECMWF systems.

3. Results

3.1. Signal-to-noise biases at the surface

To characterise signal-to-noise biases in the ECMWF forecast systems, we first make MLE estimates of model parameters for DJF hindcasts of the AO, the same data used to construct Fig. 1. Estimates of several key parameters from the model are shown in Fig. 2.

Fig. 2(a) shows the bootstrap estimate of the correlation. For both systems, there is high confidence that the correlation between the hindcasts and observations is positive (in both cases, $\Pr(\text{Correlation} < 0) < 0.01$). However, as shown by both Siegert *et al.* (2016) and Weisheimer *et al.* (2017) with a small sample size, the sampling distribution of the correlation is very broad, and it is not possible to say with confidence that the correlation of SYS4 is greater than that of SEAS5 for DJF forecasts of the AO. It is also possible to be confident that the signal-to-noise ratio of each system is too small. Fig. 2(b) illustrates this point by showing the distribution of the ratio between the signal-to-noise ratio estimated for each system and for the observations. Values less than one for this diagnostic mean that the signal-to-noise ratio in the hindcasts is smaller than the observations. There is high confidence that the signal-to-noise ratio is too

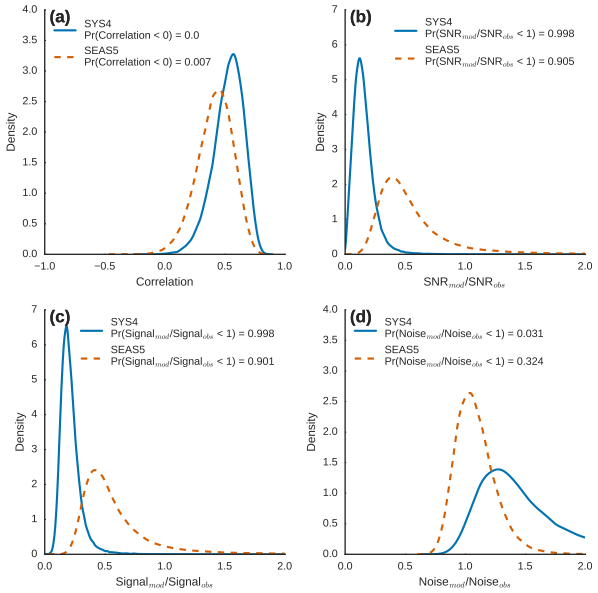

Figure 2. Bootstrap sampling distributions of parameters of the MLE statistical model for hindcasts of DJF initialised on 1st November. The blue,solid lines show SYS4 and red, dashed lines show SEAS5. Distributions are estimated using a kernel density estimator from the 10^5 bootstrap samples of each parameter. (a) shows the correlation (ρ), (b) shows the ratio of the signal-to-noise ratio in the hindcasts and observations (SNR$_x$/SNR$_y$), (c) show the ratio of the signal amplitude in the hindcasts and observations ($\beta_x/\beta y$) and (d) shows the ratio of the noise amplitude in the hindcasts and observations (η/ϵ).

low in SYS4 (Pr(SNR$_x$/SNR$_y$ $<$ 1) = 0.998) and good confidence that the signal-to-noise ratio is too low in SEAS5 (Pr(SNR$_x$/SNR$_y$ $<$ 1) = 0.905).

To further understand the origin of the low signal-to-noise ratio in each system, the bottom two panels of Fig. 2 consider the signal and noise amplitudes separately. In both systems, there is good confidence that the signal amplitude is too weak, as shown in Fig. 2(c). Additionally, for SYS4, there is high confidence that the noise amplitude is too large (Fig. 2(d)). Therefore, in SEAS5 the low signal-to-noise ratio is primarily linked to weak signal amplitude whereas for SYS4 it is a combination of weak signal amplitude and enhanced noise amplitude. This is consistent with the analysis of Siegert *et al.* (2016) for GloSea5.

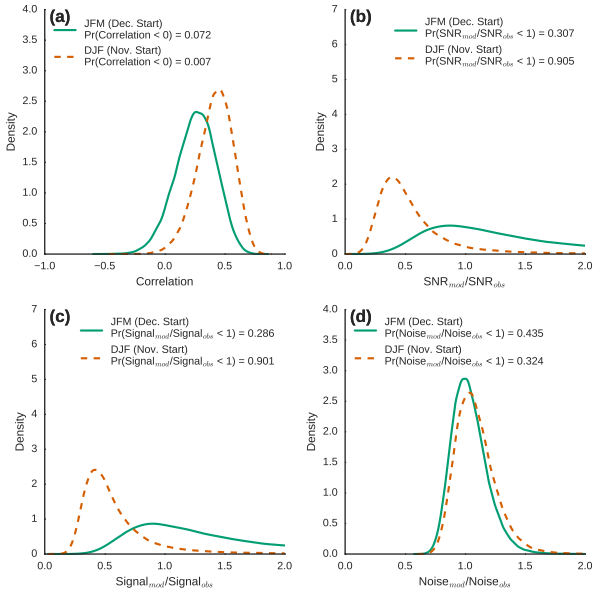

Figure 3. Bootstrap sampling distributions of parameters of the MLE statistical model for forecasts from SEAS5. JFM forecasts initialised on 1st December are shown in the green, solid line and DJF forecasts initialised on 1st November are shown in the red, dashed line (and reproduce the results shown in Fig. 2). Distributions are estimated using a kernel density estimator from the 10^5 bootstrap samples of each parameter. (a) shows the correlation (ρ), (b) shows the ratio of the signal-to-noise ratio in the hindcasts and observations (SNR$_x$/SNR$_y$), (c) shows the ratio of the signal amplitude in the hindcasts and observations ($\beta_x/\beta y$) and (d) shows the ratio of the noise amplitude in the hindcasts and observations (η/ϵ).

However, for SEAS5, the low signal-to-noise ratio present in forecasts initialised in November is not present when the system is initialised in December. Fig. 3 shows the same diagnostics as Fig. 2 but calculated for forecasts of JFM initialised on 1st December and

reproduces the diagnostics for the 1st November hindcasts of SEAS5 shown in Fig. 2. For this case, correlation between the hindcasts and observations is weaker than for November forecasts of DJF, but there is still good confidence that the correlation is positive. In contrast to the November forecasts, the signal-to-noise ratio, signal amplitude and noise amplitude of the forecast system are all consistent with estimates from the observations. Note, also, that the smaller ensemble size of the 1st December hindcasts that were analysed results in a large spread on the estimates of signal and noise in Fig. 3.

3.2. *Signal-to-noise biases throughout the atmosphere*

To further characterise the signal-to-noise biases in the forecast systems throughout the atmosphere, the same MLE estimates of the model parameters in Eq. 2 and 3 are made for the NAM index on pressure levels from 1000hPa to 10hPa. Estimates for the November forecasts for SYS4 are shown in Fig. 4 for SYS4 and for SEAS5 in Fig. 5.

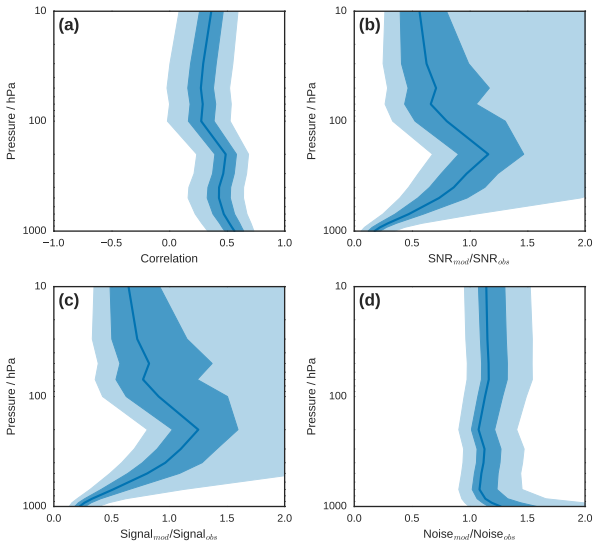

Figure 4. Bootstrap sampling distributions of parameters of the MLE statistical model for SYS4 forecasts of DJF initialised on 1st November. In each panel, the median estimate is shown by the solid line, the inter-quartile range (25th to 75th percentile) of the 10^5 bootstrap estimates is shown in dark shading and the 5th to 95th percentile is shown in the light shading. In all cases, estimates are plotted as a function of pressure level. (a) shows the correlation (ρ), (b) shows the ratio of the signal-to-noise ratio in the hindcasts and observations (SNR$_x$/SNR$_y$), (c) shows the ratio of the signal amplitude in the hindcasts and observations ($\beta_x/\beta y$), (d) shows the ratio of the noise amplitude in the hindcasts and observations (η/ϵ).

For SYS4, the low signal-to-noise ratio bias is most pronounced at the surface and decays throughout the troposphere (Fig. 4(b)) so that for pressure levels close to the tropopause there is no obvious difference between the signal-to-noise ratio in the hindcasts and the observations. As in the analysis for the surface AO, the low signal-to-noise ratio in the troposphere is the result of a combination of weak signal amplitude (Fig. 4(c)) and slightly enhanced noise amplitude (Fig. 4(d)). In the stratosphere, median estimates of the signal-to-noise ratio and signal amplitude are slightly weaker than the observations, but there is no clear and significant bias.

For SEAS5 (Fig. 5), a similar pattern to SYS4 is present. Signal-to-noise biases are strongest at the surface and through the lower and mid-troposphere and largely absent at and above the tropopause. As in the analysis of SEAS5 at the surface, the noise amplitude throughout the atmosphere is broadly consistent with the observations.

3.3. *Development of the signal-to-noise ratio bias in time*

The two previous sections show that both SYS4 and SEAS5 suffer from the widely reported low signal-to-noise ratio when skilfully forecasting the winter mean AO. This signal-to-noise bias, however, is only present for forecasts initialised on 1st November and is limited to the surface and lower and mid-troposphere. To further understand the development of this bias, it is necessary to examine how the low signal bias (the major cause of the low signal-to-noise bias in both systems) develops over the course of winter. To examine this

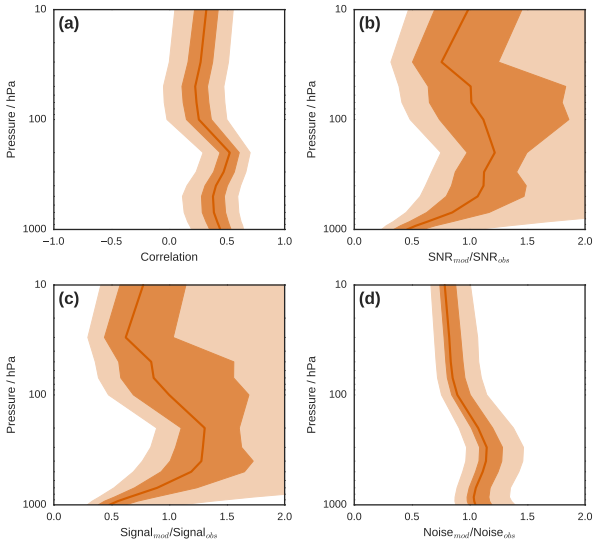

Figure 5. Bootstrap sampling distributions of parameters of the MLE statistical model for SEAS5 forecasts of DJF initialised on 1st November. In each panel, the median estimate is shown by the solid line, the inter-quartile range (25th to 75th percentile) of the 10^5 bootstrap estimates is shown in dark shading and the 5th to 95th percentile is shown in the light shading. In all cases, estimates are plotted as a function of pressure level. (a) shows the correlation (ρ), (b) shows the ratio of the signal-to-noise ratio in the hindcasts and observations (SNR$_x$/SNR$_y$), (c) shows the ratio of the signal amplitude in the hindcasts and observations ($\beta_x/\beta y$), (d) shows the ratio of the noise amplitude in the hindcasts and observations (η/ϵ).

development, MLE estimates of parameters of Eq. 2 and 3 are calculated for 30-day averages of the AO and NAM index. To determine how the low signal amplitude bias evolves in time, this procedure is repeated for each 30-day period, beginning at the initialisation time of each hindcast. For example for the 1st November forecasts 120 parameter estimates are produced beginning with the 1st to 30th November and ending with the 28th February to 29th March.

3.3.1. Surface signal

The development of the signal amplitude for the AO is shown in Fig. 6. The two sets of November 1st forecasts result in a very different development of the signal amplitude. In SYS4, (Fig. 6(a)) the signal amplitude drops rapidly from forecast initialisation so that through December, January and February the amplitude is very small, typically less than 0.3. For SEAS5, (Fig. 6(b)) the amplitude of the signal is similarly small during early winter, but increases appreciably for the late winter, particularly for forecasts covering February, as seen in other forecast systems (e.g. Jia *et al.* 2017). During this period, the signal amplitude in the hindcasts and observations is similar and the signal-to-noise ratio in the hindcasts and observations is comparable (not shown).

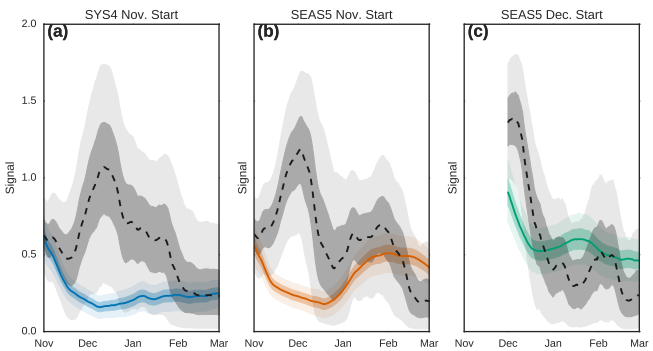

Figure 6. Bootstrap estimates of the signal term for the (a) 1st November forecasts from SYS4, (b) 1st November forecasts from SEAS5 and (c) 1st December forecasts from SEAS5. Estimates of the signal amplitude of the observations are shown in the black dashed line and grey shading, estimates of the signal amplitude for the hindcasts are shown in the solid coloured line and coloured shading. In both cases, the inter-quartile range (25th to 75th percentile) of the 10^5 bootstrap estimates is shown in dark shading and the 2.5th to 97.5th percentile is shown in the light shading.

The clearest divergence between the signal amplitude in the hindcasts and observations is during early winter for both systems. During December, the estimated signal amplitude for the observations is very large in the analysis, exceeding 1 in both cases, with either SYS4 or SEAS5. This suggests that for this set of years, there is a large, predictable signal present in the real climate system, that neither system is able to adequately capture. A similarly large amplitude signal in the observations during early winter is also present in early winter for the SEAS5 forecasts initialised in December and January (Fig. 6(c)). As in the analysis of Fig. 3, the smaller ensemble size of the analysed 1st December hindcasts likely influences the spread of the hindcast signal estimate. The anomalously low signal-to-noise ratio present in both SYS4 and SEAS5 is a result of the mismatch of signal amplitude during early winter in both systems.

3.3.2. *Development of early winter signal*

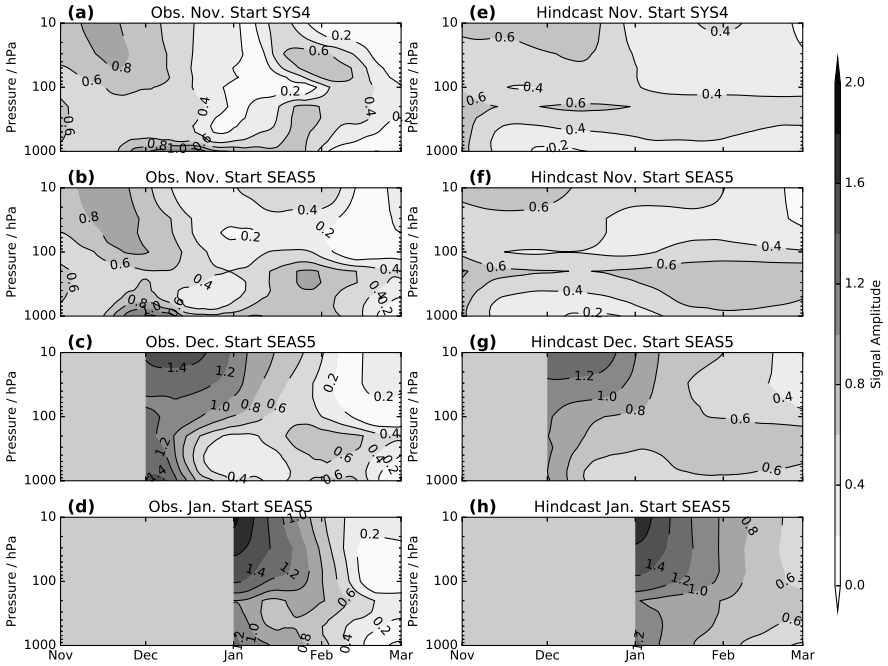

Figure 7. Maximum likelihood estimate of the signal amplitude of the observations (β_y) (a)-(d) and hindcasts (β_x) (e)-(h) on different pressure levels for 30-day periods beginning at the date plotted on the x-axis. The observational estimate differs depending on the set of hindcasts used in each calculation. All calculations are made for the NAM index. (a) and (e) show estimates for hindcasts initialised on 1st November for SYS4 and (b) and (f) show 1st November forecasts for SEAS5 (c) and (g) show 1st December forecasts for SEAS5 and (d) and (h) show 1st January forecasts for SEAS5.

To further investigate the origin of the mismatch in signal amplitude in early winter, the analysis in the previous section is repeated for other levels in the atmosphere using the NAM diagnostic. The diagnosed amplitude of the signal term in the observations is shown in Fig. 7 (a)-(d). It is immediately obvious that, as in the previous section, the structure of the predictable signal is different in early and late winter. In both SYS4 and SEAS5, the large signal amplitude at the surface in early winter is largely confined to the surface with much smaller signal amplitude through the majority of the troposphere. In contrast, in late winter, there is a second peak of large signal amplitude throughout the troposphere for monthly periods during February. A similar second peak of signal amplitude in late winter is also present for the SEAS5 forecasts initialised in December (Fig. 7(c) and (d)).

Comparing the observed signal amplitude with the signal amplitude in the hindcasts (Fig. 7 (e)-(h)) shows that the two periods in which there is significant signal amplitude in the observations are captured differently by the two systems. In early winter, in both SYS4 and SEAS5, the signal amplitude decays rapidly from the initialisation date in the lower troposphere. By December, in both systems, the amplitude of the signal in the troposphere is very weak (typically less than 0.2). In contrast, the signal amplitude in the lower stratosphere in both systems (100-300hPa) remains large throughout the winter season. By late winter, the signal amplitude in this region is comparable in the hindcasts and observations. In SEAS5, larger signal amplitude extends through the troposphere to the

surface, consistent with the AO diagnostics in Fig. 6. In SYS4, the signal amplitude in the lower stratosphere and its extension to the surface is weaker. For SEAS5 forecasts initialised in December (Fig. 7(g)) a similar amplification of the hindcast signal occurs in late winter.

4. Discussion and Conclusions

In this study, the statistical model of Siegert *et al.* (2016) is used to explore the signal-to-noise properties of forecasts of the Northern Annular Mode and Arctic Oscillation made by two ECMWF seasonal forecasting systems. By adapting the methodology from its original Bayesian framework it is possible to explore the origin of the signal-to-noise problem (Eade *et al.* 2014) in SYS4 identified by Stockdale *et al.* (2015) and common in other systems (Baker *et al.* 2018). The analysis in the present study confirms that both SYS4 and SEAS5 suffer from anomalously low signal-to-noise ratio for DJF forecasts when initialised on November 1st. However, by fitting the Siegert *et al.* (2016) model to other pressure levels and forecasts initialised at other times of the year it is possible to refine understanding of the properties of the signal-to-noise problem, at least in the ECMWF systems, namely that:

- A significant contribution to the low signal-to-noise ratio in both systems is related to an anomalously small signal amplitude as has previously been shown for other systems (e.g. Siegert *et al.* 2016).
- In SEAS5, the low signal-to-noise ratio is present for forecasts initialised on November 1st but not for forecasts initialised on December 1st.
- The low signal-to-noise ratio is present at the surface and through the lower and mid-troposphere but not the stratosphere.
- The low signal amplitude in the forecast systems is dominated by a mismatch in signal amplitude between the hindcasts and observations in early winter.
- The low signal amplitude in early winter is predominantly a lower tropospheric feature.

Together, these conclusions suggest that in order to further understand (and ultimately correct) the low signal-to-noise problem in the ECMWF seasonal forecasting systems, further experimentation should be focussed on processes that dominate the predictability of early winter at the surface. Our results are consistent with the statement in Stockdale *et al.* (2015) that in SYS4 "The weakness of the stratosphere to troposphere coupling in the model results in a weak signal at the surface and a major underestimate of predictability" and show that this problem is somewhat ameliorated in SEAS5. It would be extremely interesting, but beyond the scope of the present study, to repeat some of the analysis with other well-studied seasonal forecasting systems such as GloSea5. In particular, it is important to establish if the low SNR is confined to DJF forecasts and the lower troposphere as it is in SEAS5.

The period of apparently large predictable signal during early winter is particularly interesting and also deserves further study. The characteristics and predictability of early winter circulation may be affected by longer term variability. As noted by Hanna *et al.* (2015), during recent winters there has been a large increase in the variability of the NAO and analysis of hindcasts initialized over a 110-year period (Weisheimer *et al.* 2018) suggests that there is significant variation of over and under confidence during the 20th century linked to long-term variability of the global circulation and teleconnections (e.g. O'Reilly 2018). Analysis of a much longer set of seasonal forecasts such as that described in Weisheimer *et al.* (2017) would be a useful first step in exploring differences in predictable signals in early and late winter. Weisheimer *et al.* (2018) also see differences in over and under confidence when they are measured with different diagnostics, and suggest that when the predictable signal in the hindcasts is too weak, some diagnostics of over and under confidence are more sensitive to a change in the amplitude of the signal than others. It is important to note, though, that the results of Weisheimer *et al.* (2018) might not be applicable to the kind of analysis proposed for two reasons: 1) because the initialization contains little stratospheric information prior to 1979, which is now known to be important for NAO prediction skill, making comparison of different periods is misleading and 2) the simulations are not seasonal forecasts but rather atmosphere only simulations.

Analysis of forecasts initialized in December, when the hindcast signal is larger, could shed further light on how different measures of over and under confidence compare. As noted by other authors, estimates of many of the parameters in our statistical model are very uncertain given the short hindcast record. We have attempted to exercise caution in interpreting differences in estimates of, for example, the signal term in the observations since these estimates are sensitive to the modelling system used in the analysis.

There are a number of possible reasons why the ECMWF forecast systems might be unable to capture the available skill in early winter. O'Reilly *et al.* (2018) showed evidence that the link between the Quasi-Biennial Oscillation and the NAO was too weak in a version of the ECMWF model. QBO effects in the polar stratosphere, while often hard to untangle from other drivers of variability, are thought to be most significant from October to December (Garfinkel and Hartmann 2007). Shaw *et al.* (2010) show that wave reflection in the stratosphere has two distinct peaks, one in early winter and one in late winter, but they also show that wave coupling between the stratosphere and troposphere is not significantly enhanced during early winter in the years studied (1979-2001). The extent to which the ECMWF forecast system is able to capture wave reflection in early winter, and if this reflection played a role in the predictable signal, remains to be studied.

Acknowledgement

This work was completed while AJCP was on sabbatical at ECMWF during 2017 and 2018. He would like to thank ECMWF and particularly Magdalena Balmaseda for their hospitality during this time.

A. Deriving contrasts

The following algorithm is used to create a set $\{\mathbf{w}^{(m)} \in \mathbb{R}^K, m = 1, \ldots \mu\}$ of contrasts, where necessarily $\mu < K$.

1. Let V be a matrix of dimension $K \times (\mu + 1)$ with rank $(\mu + 1)$ (i.e. the columns are linearly independent) and the first column being a constant vector (i.e. all entries are the same and not zero). For instance we could take:

$$\mathsf{V}_{k,l} = \left(\frac{k}{K+1} - \frac{1}{2}\right)^{l-1}$$

2. Let Q, R be matrices of dimension $K \times (\mu + 1)$ and $(\mu + 1) \times (\mu + 1)$, respectively, so that
   (a) the columns of Q are normalised and mutually orthogonal;
   (b) R is right upper triangular;
   (c) $\mathsf{V} = \mathsf{QR}$.

   Such matrices can be found by applying a Gram–Schmidt procedure to the columns of V or equivalently through a QR–decomposition of V.
3. Now ignore the first column of Q which will have constant entries; the remaining μ columns form the desired contrasts.

References

Baker L, Shaffrey L, Sutton R, Weisheimer A, Scaife 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.

Baldwin MP, Thompson DW. 2009. A critical comparison of stratosphere–troposphere coupling indices. *Quarterly Journal of the Royal Meteorological Society* **135**(644): 1661–1672.

Butler AH, Arribas A, Athanassiadou M, Baehr J, Calvo N, Charlton-Perez A, Déqué M, Domeisen DI, Fröhlich K, Hendon H, *et al.* 2016. The Climate-system Historical Forecast Project: do stratosphere-resolving models make better seasonal climate predictions in boreal winter? *Quarterly Journal of the Royal Meteorological Society* **142**(696): 1413–1427.

Dee DP, Uppala S, Simmons A, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda M, Balsamo G, Bauer dP, *et al.* 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society* **137**(656): 553–597.

Domeisen DI, Butler AH, Fröhlich K, Bittner M, Müller WA, Baehr J. 2015. Seasonal predictability over Europe arising from El Niño and stratospheric variability in the MPI-ESM seasonal prediction system. *Journal of Climate* **28**(1): 256–271.

Dunstone N, Smith D, Scaife A, Hermanson L, Eade R, Robinson N, Andrews M, Knight J. 2016. Skilful predictions of the winter North Atlantic Oscillation one year ahead. *Nature Geoscience* **9**(11): 809.

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.

Garfinkel CI, Hartmann DL. 2007. Effects of the El Niño–Southern Oscillation and the quasi-biennial oscillation on polar temperatures in the stratosphere. *Journal of Geophysical Research: Atmospheres* **112**(D19).

Hanna E, Cropper TE, Jones PD, Scaife AA, Allan R. 2015. Recent seasonal asymmetric changes in the NAO (a marked summer decline and increased winter variability) and associated changes in the AO and Greenland Blocking Index. *International Journal of Climatology* **35**(9): 2540–2554.

Jia L, Yang X, Vecchi G, Gudgel R, Delworth T, Fueglistaler S, Lin P, Scaife AA, Underwood S, Lin SJ. 2017. Seasonal prediction skill of northern extratropical surface temperature driven by the stratosphere. *Journal of Climate* **30**(12): 4463–4475.

Johnson SJ, Stockdale TN, Ferranti L, Balmaseda MA, Molteni F, Magnusson L, Tietsche S, Decremer D, Weisheimer A, Balsamo G, Keeley S, Mogensen K, Zuo H, Monge-Sanz B. 2018. SEAS5: The new ECMWF seasonal forecast system. *Geosci. Model Dev. Discuss.* **in review**.

Kidston J, Scaife AA, Hardiman SC, Mitchell DM, Butchart N, Baldwin MP, Gray LJ. 2015. Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. *Nature Geoscience* **8**(6): 433.

Molteni F, Stockdale T, Balmaseda M, Balsamo G, Buizza R, Ferranti L, Magnusson L, Mogensen K, Palmer T, Vitart F. 2011. *The new ECMWF seasonal forecast system (System 4)*. European Centre for Medium-Range Weather Forecasts Reading, U. K.

O'Reilly CH. 2018. Interdecadal variability of the ENSO teleconnection to the wintertime North Pacific. *Climate Dynamics* : 1–18.

O'Reilly CH, Weisheimer A, Woollings T, Gray L, MacLeod D. 2018. 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* .

Scaife A, Arribas A, Blockley E, Brookshaw A, Clark R, Dunstone N, Eade R, Fereday D, Folland C, Gordon M, *et al.* 2014. Skillful long-range prediction of European and North American winters. *Geophysical Research Letters* **41**(7): 2514–2519.

Scaife A, Karpechko AY, Baldwin M, Brookshaw A, Butler A, Eade R, Gordon M, MacLachlan C, Martin N, Dunstone N, *et al.* 2016. Seasonal winter forecasts and the stratosphere. *Atmospheric Science Letters* **17**(1): 51–56.

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

Schervish MJ. 1995. *Theory of statistics*. Springer Series in Statistics, Springer-Verlag, New York, ISBN 0-387-94546-6, doi:10.1007/978-1-4612-4250-5.

Shaw TA, Perlwitz J, Harnik N. 2010. Downward wave coupling between the stratosphere and troposphere: The importance of meridional wave guiding and comparison with zonal-mean coupling. *Journal of Climate* **23**(23): 6365–6381.

Siegert S, Stephenson DB, Sansom PG, Scaife AA, 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.

Stockdale TN, Molteni F, Ferranti L. 2015. Atmospheric initial conditions and the predictability of the Arctic Oscillation. *Geophysical Research Letters* **42**(4): 1173–1179.

Weisheimer A, Decremer D, MacLeod D, O'Reilly C, Stockdale T, Johnson S, Palmer T. 2018. How confident are predictability estimates of the winter North Atlantic Oscillation? *Quarterly Journal of the Royal Meteorological Society* .

Weisheimer A, Schaller N, O'Reilly C, MacLeod DA, 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.

Zuo H, Balmaseda MA, Tietsche S, Mogensen K, Mayer M. 2019. The ECMWF operational ensemble reanalysis-analysis system for ocean and sea-ice: a description of the system and assessment. *Ocean Sci. Discuss.* .