



Signal and noise in regime systems: a hypothesis on the predictability of the North Atlantic Oscillation

K. Strommen*, T.N. Palmer

Oxford University, Clarendon Laboratory, AOPP, Parks Road, OX1 3PU, Oxford.

*Correspondence to: Oxford University, Clarendon Laboratory, AOPP, Parks Road, OX1 3PU, Oxford. E-mail: kristian.strommen@physics.ox.ac.uk

Studies conducted by the UK Met Office (Dunstone *et al.* (2016), Eade *et al.* (2014), Scaife *et al.* (2014)) reported significant skill at predicting the winter NAO index with their seasonal prediction system. At the same time, a very low signal-to-noise ratio was observed, as measured using the ‘ratio of predictable components’ (RPC) metric. We analyse both the skill and signal-to-noise ratio using a new statistical toy-model which assumes NAO predictability is driven by regime dynamics. It is shown that if the system is approximately bimodal in nature, with the model consistently underestimating the level of regime persistence each season, then both the high skill and high RPC value of the Met Office hindcasts can easily be reproduced. Underestimation of regime persistence could be attributable to any number of sources of model error, including imperfect regime structure or errors in the propagation of teleconnections. In particular, a high RPC value for a seasonal mean prediction may be expected even if the models internal level of noise is realistic.

Key Words: winter predictability, atmosphere, seasonal prediction, regimes, NAO, persistence, signal-to-noise ratio

Received ...

1. Introduction

The North Atlantic Oscillation (NAO), being the dominant mode of variability in the North Atlantic, plays a critical role in modulating European and North-American climate, particularly in the winter months (e.g. Marshall *et al.* (2001), Greatbatch (2000)). For many years, seasonal prediction forecasts produced using general circulation models (GCM’s) showed no significant skill at predicting the wintertime NAO, and it was widely thought that this was an inherently chaotic process with no predictability on seasonal timescales (Johansson (2007)). In more recent years, this view has been challenged by various studies: see e.g. Smith *et al.* (2016) for an overview. In Müller *et al.* (2005), some models were shown to exhibit skillful wintertime predictions of the NAO at a lead-time of 1 month, though the skill appeared limited to the shorter time-period 1987-1999; none of the models showed noteworthy skill when extended to a longer hindcast period. More recently, studies conducted using the Met Office Global Seasonal Prediction System 5 (GloSea5) displayed skillful ensemble hindcasts of the winter NAO (Scaife *et al.* (2014)) over a 20 year hindcast period. This skill was then corroborated by studies conducted with the UK Met Office Decadal Prediction System 3 (DePreSys3) in Dunstone *et al.* (2016), which showed that this skill remained when assessed over a longer hindcast period of 35 years. Bayesian analysis conducted in Siegert *et al.* (2016) indicated that this skill was unlikely to be a result of chance. A multimodel study carried out in Athanasiadis *et al.* (2017) showed this high skill was attainable by other models as well.

At the same time, it was observed that the model seemed to have a notably low signal-to-noise ratio (Scaife *et al.* (2014)). Indeed, the skill of 0.62 found in Dunstone *et al.* (2016), measured as the correlation between the ensemble mean of the hindcast NAO index and the NAO index as observed in re-analysis, could only be achieved by taking the mean of a large ensemble. On the other hand, the skill of the model at predicting itself (measured as the average correlation between the ensemble mean and individual members), was significantly lower, at 0.18. That is, the model NAO appears to be notably less predictable than the real NAO. This was referred to as the ‘signal-to-noise paradox’ in Dunstone *et al.* (2016). In Eade *et al.* (2014) and Dunstone *et al.* (2016) this ‘paradox’ was encapsulated using the ‘ratio-of-predictable-component’ (RPC) metric, introduced in Eade *et al.* (2014). In an ideal model which captures perfectly the level of predictability of the true climate system, this metric will equal 1. An RPC less than 1 indicates that the models internal NAO is too predictable, while an RPC greater than 1 indicates it is not predictable enough. This metric, based on correlations, is subtly different from the perhaps more familiar ‘root-mean-square error divided by spread’ metrics commonly used to diagnose ensemble skill relative to noise (see e.g. Fortin *et al.* (2014) and references within). In particular, by being insensitive to magnitudes, the RPC is able to more sensitively examine situations where the signals are particularly small. The ‘signal-to-noise paradox’ can be summarized by the

observation that the *RPC* in the DePreSys3 hindcasts is 2.31, significantly greater than 1. The fact that the estimate from the original 20 year hindcast considered in Eade *et al.* (2014) remained robust when extended to the 35 year hindcast considered in Dunstone *et al.* (2016) suggests that this estimate is not purely a result of sampling variability. While Shi *et al.* (2015) pointed out the sensitivity of the *RPC* estimate to even longer hindcast lengths, suggesting that the *RPC* may be subject to decadal variability, it never the less left open the question as to why the *RPC* is notably large in the last part of the century. Furthermore, given the great uncertainty inherent in estimates of the *RPC* in situations with low model skill (see section 2.2 and 3.4), a point not emphasized in Shi *et al.* (2015), it is likely that more work is required to understand such decadal fluctuations.

All of this raises an obvious question: can we characterize the nature of the model skill required to produce this behaviour? In particular, is a high *RPC* value expected or not given the dynamics of the underlying system? The answer to this question involves the construction (or choice) of a statistical null-hypothesis, in other words a toy-model of the NAO. For example, in Scaife *et al.* (2014), the work conducted in Kumar (2009), which used a simple linear shift toy-model, was cited as evidence that the observed skill was higher than what one would expect given the model's level of noise, implying that the observed *RPC* is spuriously high. In Siegert *et al.* (2016), a 'signal+noise' toy-model was used to assess the statistical significance of the observed behaviour using Bayesian techniques, concluding that both the high skill and the low signal-to-noise ratio were significant. Both these toy-models are essentially variants of the same 'linear-regression paradigm'. In this paradigm of predictability, ensemble members are viewed as being normally distributed noise around an ensemble mean which detects, to some level of accuracy, a predictable signal in the *magnitude* of the NAO.

Many studies have argued that the NAO exhibits regime behaviour (see e.g. Straus *et al.* (2007), Woollings *et al.* (2010), Corti *et al.* (1999)). The regime paradigm offers another potential type of predictability, namely predictability of preferred regime states. The potential importance of shifts in preferred atmospheric states for predicting the long-term climate response to anthropogenic forcing was explained in Palmer (1999), but this mechanism can be considered also on seasonal timescales. Over a given winter season, one can envisage the atmosphere, on each day, falling into a particular regime state. If the model's atmosphere has a preference towards spending more days in regimes that induce positive NAO indices (for example due to some external forcing), one would expect the seasonal mean to be positive, and similarly for negative states. If the preferred model state coincides with the preferred state in reality, this will induce positive correlation. In this way, one can envisage that in a given season it may be possible to have predictability of the *sign* of the NAO index which is independent of any predictability of the *magnitude*. Indeed, it may be that in some seasons, the sign is essentially the *only* thing predictable.

In this paper we will analyse a new toy-model aimed at capturing such regime behaviour. In this toy-model, the NAO is explicitly represented as an idealized 2-state Markov chain, with seasonally varying transition probabilities. Error is introduced into ensemble members by modulating the extent to which they are able to capture these probabilities. Specifically, we will assume that, ignoring noise from internal variability, the ensemble members all underestimate the persistence probabilities of both regimes. We show that this toy-model can easily capture the observed high-*RPC* behaviour, even when the internal level of noise in the model perfectly matches that of observations: the apparently excessive noise on seasonal timescales is entirely a consequence of model error on daily timescales. Concretely, the combination of high ensemble mean correlations and high *RPC* values is to be expected in this toy-model due to the weak persistence alone, and as such, may be attributable to any number of model errors, including e.g. imperfect regime structure or errors in the propagation of teleconnections. In particular, low signal-to-noise ratios can be expected due to model bias even when the model's internal variability (i.e. noise) is perfectly realistic.

Finally, we compare the *RPC* metric to another common way to measure if a model is under/over-confident, namely the ratio of the root-mean-square error of the ensemble mean prediction with the ensemble spread. It is shown that, unless the GCM is statistically perfect, these two metrics in general may give entirely different answers.

Because the skill observed in Dunstone *et al.* (2016) with DePreSys3 is so high over a long 35-year window, we will be focusing on replicating the explicit numbers from that study in this paper. While the results will therefore be aiming at reproducing this particular example of the 'signal-to-noise paradox', we note that we believe the model presented could equally well be fitted to data from other GCMs with comparable results.

2. Methodology and definitions

2.1. Definition of relevant metrics

We will primarily be interested in three different metrics assessing model skill, all closely related. We define $Corr(EnsMean, Obs)$ to be the Pearson correlation coefficient between the ensemble mean of some forecast and observations over a given time-period. We will refer to this informally as the *actual predictability*. On the other hand, we define $Corr(EnsMean, Mem)$ to be the average correlation between the ensemble mean and individual members of the ensemble. This is sometimes referred to as the 'potential predictability', but we will avoid this term as it is misleading. Indeed, as we shall see, this quantity can often be significantly lower than the actual predictability, and therefore does not provide a well-defined upper bound on the level of predictability.* Instead, we will refer to this quantity as the *model predictability* of the system, as it measures the model's ability to predict itself.

The third metric of interest is the 'ratio of predictable components', *RPC*, defined in Eade *et al.* (2014). One first defines the notion of the 'predictable component' of a given system as the square root of the total fraction of variance that can be predicted in that system. When applied to observations we obtain $PC(Obs)$, and when applied to a model we obtain $PC(Mod)$. The *RPC* is then by definition

$$RPC_{true} = PC(Obs)/PC(Mod) \quad (1)$$

In Eade *et al.* (2014) they provide an estimate of *RPC* amenable to computation via the following inequality:

$$RPC_{true} = \frac{PC(Obs)}{PC(Mod)} \geq \frac{\sqrt{Corr(EnsMean, Obs)^2}}{\sqrt{\sigma_{sig}^2/\sigma_{tot}^2}} \quad (2)$$

*This observation was also noted in Kumar *et al.* (2014).

94 where σ_{sig}^2 is the ensemble mean variance and σ_{tot}^2 is the average variance of individual ensemble members. We define the RPC
95 estimate to be this last term:

$$RPC_{est} = \frac{Corr(EnsMean, Obs)}{\sqrt{\sigma_{sig}^2 / \sigma_{tot}^2}} \quad (3)$$

96 It is important to note that this lower bound has been obtained by choosing a positive square root, and so is always positive, even if
97 the correlation between ensemble mean and observations could, in principle, be negative. Thus our estimate of the *RPC* is in fact the
98 absolute value of this last term.[†]

99 For a model to be ‘statistically perfect’, one firstly requires that different ensemble members are statistically indistinguishable
100 from each other, and therefore that they share the same variance: $\sigma_{tot} = \sigma_i$ for all i , where σ_i is the standard deviation of ensemble
101 member i . Given this assumption, it is easy to show that the expected value of the denominator of the *RPC* estimate above is exactly
102 equal to the model predictability, and so RPC_{est} equals the ratio of actual to model predictability. Secondly, for a ‘perfect’ model,
103 observations are required to be statistically indistinguishable from individual members (a property also known as ‘exchangability’).
104 Under this assumption, it is easy to show that, in expectation, the internal and actual predictability will coincide. If we define ‘statistical
105 perfection’ to be met when both these conditions are met, it follows that for a perfect model, the *RPC* will equal 1. An $RPC > 1$
106 corresponds to the model being more skillful at predicting observations than itself, with the opposite being true when $RPC < 1$.

107 In practice, in the cases considered in our analysis, the assumption that different ensemble members share the same variance will not
108 typically be true due to sampling variability and the fact that the ensembles considered will only have 40 members. Therefore, when
109 computing it we always use the explicit estimate. However, equating *RPC* with the ratio of actual to model predictability is useful for
110 guiding intuition and interpretation.

111 2.2. A cautionary note on computing the *RPC*

112 The estimate RPC_{est} is obtained by using the ensemble mean as a proxy for reality, so one can compute $PC(Obs)$. If the model is very
113 skillful, then even a small ensemble will produce an ensemble mean which is a good approximation to reality. However, in situations
114 where the inherent model skill is low, this will no longer be the case. In fact, given a fixed ensemble size, the estimate breaks down
115 dramatically as model skill decreases.

116 Indeed, if the model has no skill whatsoever, then both actual and model predictability are expected to be 0. In particular,
117 $PC(Mod) = 0$ and the ensemble members and observations are all uncorrelated. However, assuming that the true system has *some*
118 predictable component, $PC(Obs) > 0$, and so $RPC_{true} = \infty$. On the other hand, if we assume that in this ‘no-skill’ situation, all
119 the ensemble members and reality are behaving as independent normal distributions, we can approximate all the terms in equation
120 (3). If we assume that each individual ensemble member follows a distribution with variance σ_m^2 , and that there are enough ensemble
121 members to make $\sigma_{tot} = \sigma_m$ a decent approximation, then the denominator of (3) will be $1/\sqrt{N}$. To estimate the numerator, first note
122 that the distribution of correlation coefficients between independent normal distributions follows a Student’s T-distribution, which,
123 given a sufficiently large sample size K can be approximated as normal with mean 0 and variance $1/\sqrt{K}$. Since we are taking absolute
124 values, we also need to note that if Y is a normal distribution with mean 0 and variance σ^2 , then $|Y|$ has mean $\sigma\sqrt{2/\pi}$. Putting all this
125 together, we find that when the model has no skill,

$$RPC_{est} \approx \sqrt{2N/\pi K} \quad (4)$$

126 where N is the ensemble size and K is the length of the timeseries being correlated.[‡] For example, with $N = 40$ and $K = 35$, as in
127 Dunstone *et al.* (2016), this is approximately 0.85. If one does not take absolute values in RPC_{est} , the expected *RPC* estimate can
128 similarly be shown to be exactly 0. So if the model has little to no skill, and the ensemble size N has been fixed, the true *RPC* will
129 diverge to ∞ , while the estimate could be close to 0 or 1 depending on how it is computed! On the other hand, if we leave model skill
130 fixed (e.g. by fixing a model cycle) then since the expression (4) diverges to ∞ as N grows, RPC_{est} can be made into an arbitrarily
131 better estimate of RPC_{true} by increasing the ensemble size.

132 The moral of the story is that:

- 133 (1) It is important to be clear about exactly how one has computed the *RPC* estimate, in particular whether absolute values have
134 been taken or not.
- 135 (2) In situations where the model skill or ensemble size is small, great caution must be taken when interpreting estimates of
136 RPC_{true} . Depending on the method used, an RPC_{est} close to 0 or 1 may be indicative of nothing more than the fact that
137 the actual predictability is low, and cannot be used to infer anything about the true *RPC*.

138 This is of particular importance when e.g. computing *RPC* over time-periods where the model exhibits little skill, where a typical,
139 modern ensemble size of around 40-50 may be insufficient to robustly estimate it.

140 2.3. Data

141 The DePreSys3 ensemble data was kindly provided to us by Nick Dunstone, and is what we will be comparing our toy-model against.
142 When the observational NAO index is referred to, we are using the same observational dataset considered in Dunstone *et al.* (2016),
143 namely HadSLP (see Allan and Ansell (2006)). This data was also generously provided to us by Nick Dunstone. In this way, we can
144 faithfully compare our toy-models to the results of Dunstone *et al.* (2016). In their paper, the NAO was calculated as the difference in
145 mean sea level pressure averaged over boxes around Iceland and the Azores.

[†]Note that in Siebert *et al.* (2016) this estimate (without the absolute value) is taken as the *definition* of *RPC*. We will stick with the original definition in this paper.

[‡]We will see later that this approximation is very good in the situation we are in.

146 The determination of various regime parameters in section 3.1 was done using the reanalysis product ERA-Interim (Dee *et al.*
147 (2011)). As an example of models lack of persistence, we used data from GloSea5[§] (see ?), available freely via the Copernicus
148 Project (Raoult *et al.* (2017)). This dataset consisted of 21 ensemble members covering the period 1993-11-01 to 2016-03-01, thereby
149 comprising 483 years of model data.

150 3. The Markov model

151 3.1. The NAO index as a 2-state Markov chain

152 The signal+noise model of Siegert *et al.* (2016) assumed that the seasonally varying probability distribution function (pdf) of possible
153 NAO indices is normally distributed with a mean which is shifted linearly from season to season according to a predictable scaling
154 factor. Model skill is accordingly induced by the extent to which the model correctly detects this linear shift. In this section we introduce
155 our new Markov model, aiming to capture the predictability expected in a more non-linear, regime based system. **We will aim to show**
156 **that such a model provides another hypothetical explanation for the ‘signal-to-noise paradox’.**

157 From a regime perspective, a season is often thought of as being built up of daily events (e.g. (Dawson *et al.* 2012)), a convention
158 we will follow. A DJF season thereby consists of approximately 90 days, where for each day, the atmosphere resides in one of the
159 available regimes. In this way, a 90 day mean is generated by sampling a Markov process over 90 steps, and as such is determined both
160 by the magnitude of individual events and the matrix of transition probabilities. If the discrepancy between the magnitudes of events in
161 different states is small, the long-term mean will be determined almost uniquely by the transition probabilities alone.

162 This can be made explicit in simple situations. Figure 1 shows a schemata for a basic, 2-state Markov chain with transition
163 probabilities α and β (ranging between 0 and 1) and two states labelled - and +, corresponding to the NAO- and NAO+. Suppose
164 that each timestep in this chain corresponds to one day, and that each day the atmosphere ends up in the minus state, the value of the
165 NAO index on that day is a random draw from the probability distribution X_- . Conversely, on days where it ends up in the plus state,
166 the NAO index is a random draw from X_+ . Denote the pdf’s of both as $\rho_+(x)$ and $\rho_-(x)$ respectively.

167 Let π_- and π_+ denote the long-term average proportion of time spent in state - and + respectively. It is an exercise in basic Markov
168 theory (see Appendix A for a sketch of the derivation) to show that

$$\begin{aligned}\pi_- &= (1 - \beta)/(2 - \alpha - \beta) \\ \pi_+ &= (1 - \alpha)/(2 - \alpha - \beta)\end{aligned}$$

169 and hence the preferred state is determined by the relative values of α and β , with no preferred state if and only if $\alpha = \beta$.

170 If $Y = Y(\alpha, \beta)$ represents the distribution of N-day means obtained from this process, then for sufficiently large values of N , we
171 have

$$Y \sim \frac{1}{N} \left(\sum_{N(+)} X_+ + \sum_{N(-)} X_- \right) \quad (5)$$

$$\sim \frac{1}{N(+)} \sum_{N(+)} \pi_+ X_+ + \frac{1}{N(-)} \sum_{N(-)} \pi_- X_- \quad (6)$$

172 where $N(+)$ $\approx N\pi_+$ is the expected number of timesteps spent in the + state and $N(-)$ $\approx N\pi_-$ the expected number of timesteps
173 spent in the - state. By the Central Limit Theorem, and the fact that the sum of normal distributions is normal, Y will be approximately
174 normally distributed with expected value

$$E(Y) = \pi_+ E(X_+) + \pi_- E(X_-) \quad (7)$$

175 and variance

$$Var(Y) = \frac{\pi_+ Var(X_+) + \pi_- Var(X_-)}{N} \quad (8)$$

176 Let us further assume that X_+ is strictly positive, and that $\rho_-(-x) = \rho_+(x)$ for all $x > 0$; in other words, NAO+ days always have
177 positive index, and NAO- days always have negative index, with the distribution of each being equal up to a sign. If $\beta > \alpha$, then
178 $\pi_+ > \pi_-$. The NAO+ is now the preferred state, and the expected mean and variance of Y is

$$E(Y) = (\pi_+ - \pi_-)E(X_+) \quad (9)$$

179 and

$$Var(Y) = Var(X_+)/N \quad (10)$$

180 From equation (5) we can further deduce that, for sufficiently large N , the $N(+)$ negative contributions from X_- will approximately
181 cancel an equal number of positive contributions from X_+ , leaving on average $N(+)$ - $N(-)$ strictly positive contributions. Therefore,
182 Y is expected to be strictly positive. The situation is opposite for NAO- years when Y is expected to be strictly negative.

183 In a seasonal mean, $N = 90$. If α and β are close to each other, this may not be a sufficient number of days to witness the above
184 expected trend, even with a statistically perfect model. In particular, the pdf of seasonal means may not end up being strictly positive
185 or negative. This would correspond to a year with a low level of predictability. If the probabilities are more well separated on the other
186 hand, one would be able to detect the impact on the mean robustly over a short period of time, corresponding to a year with a high
187 level of predictability. This predictability would be induced by seasonal deviations in the persistence/transition probabilities from their

[§]Note that this is the same model as used in the DePreSys3 system.

188 climatological means, and the models ability to detect these. Such deviations could be driven by some specific external forcing (e.g.
189 tropical sea surface temperatures, the stratosphere, sea ice extent, etc. as suggested in [Dunstone *et al.* \(2016\)](#)), internal dynamics (e.g.
190 [transient baroclinic eddy feedback](#)), or a complex combination of the two.

191 In the papers [Dawson *et al.* \(2012\)](#), [Dawson and Palmer \(2015\)](#), clustering algorithms are used to argue for the existence of four
192 distinct regimes, commonly titled NAO+, NAO-, Blocking and Atlantic Ridge. However, other studies suggest there may be three
193 ([Woollings *et al.* \(2010\)](#)), or even only two regimes ([Woollings *et al.* \(2008\)](#)), suggesting the picture remains uncertain. For the sake
194 of simplicity, we will view the North Atlantic as a 2-state system, in line with [Woollings *et al.* \(2008\)](#), where positive NAO events
195 correspond to more consistently zonal, unblocked jet, while the negative NAO corresponds to more consistently blocked jet. In terms
196 of the four regime picture of [Dawson *et al.* \(2012\)](#), this amounts to amalgamating the NAO-, Blocking and Atlantic ridge regimes into
197 one big regime representing blocked flow, and leaving the NAO+ as representing zonal flow.

198 Before testing this model in the next section, we point out that this should be seen as a ‘proof-of-concept’ model. It seems unlikely
199 that the NAO could be easily captured in its entirety by such a simple 2-state process, but while one could potentially extend the analysis
200 to a multi-state system, we will see that with two states one already captures many features of the observed NAO predictability. An
201 extension of this framework to four states will be considered in future work.

202 3.2. Potential sources of model error

203 Using, as explained above, the 2-state Markov chain as a toy-model for the NAO index, any GCM aiming to predict the NAO needs to
204 accurately predict changes in the transition/persistence probabilities in each season in order to capture the predictable component. In
205 reality, models are known to suffer from systematic biases, and so errors are expected to be present in the Markov model. We propose
206 four different potential sources of model error that may be expected to interact with the regime structure:

- 207 (1) There is too much inherent stochasticity in the system. Imagining the regime states as potential wells, this corresponds to a
208 model where if the atmosphere ends up in a given well, it will quickly be pushed out of it due to the high level of noise in the
209 system, systematically warping estimates of transition probabilities.
- 210 (2) The regime structure in the model is too weak, or suffers from persistent biases. In this setting, there is no excess stochasticity
211 in the system, but the potential wells themselves may for example be too shallow, meaning that even realistic levels of noise is
212 sufficient to push the atmosphere out of a given regime too quickly.
- 213 (3) If we assume that variations in transition probabilities from year to year are driven to a large extent by specific sources of external
214 forcing (i.e. global dynamics), then it may be that the model is simply unable to correctly propagate this forcing consistently.
215 In this way, even if the model has perfect regime structure *and* realistic levels of noise, it may still fail to predict the correct
216 probabilities.
- 217 (4) If we instead assume that internal dynamics are the key driver to changes in transition probabilities, it may be that even tiny
218 errors in initial conditions are sufficient to produce ensemble members that fail to capture the observed transition probabilities,
219 even if the model is otherwise very skillful.

220 If one assumes that the North Atlantic does have regimes, then any error would necessarily filter through the regimes in a manner
221 not obvious a priori. Therefore, while only error (2) in the above list specifically pertains to deficiencies in the regimes themselves, the
222 other errors may still be important to understand in the context of our statistical model. However, precisely because only error (2) is
223 explicitly related to regime deficiencies, we will focus on this possibility in the remainder of the paper.

224 That climate models tend to have persistent biases when it comes to their representation of North Atlantic regime structure has been
225 observed in e.g. [Dawson *et al.* \(2012\)](#). Indeed, in re-analysis, persistence is typically more common than transition, with persistence
226 probabilities of all North Atlantic regimes being greater than one half. However, models typically do not obtain similar levels of
227 persistence, struggling e.g. to achieve long-lived blocking events ([Scaife *et al.* \(2010\)](#), [Davini and D’Andrea \(2016\)](#)). This implies that
228 a typical model error is to underestimate persistence probabilities. This would be consistent with both (1) and (2) above, as both might
229 be expected to reduce model persistence.

230 As a concrete example of this, we compare persistence of the NAO+ regime in ERA-Interim to that in the GloSea5 model: we
231 remind the reader that this is the same underlying model as DePreSys3 (see section 2.3). To do so, the method detailed in [Dawson
232 *et al.* \(2012\)](#) was used to compute regimes in both datasets using K-means clustering. The clustering is performed on daily, detrended
233 anomalies of geopotential height at 500 hPa over the Euro-Atlantic region in the DJF season; the dimensionality is first reduced by
234 projecting re-analysis and model data onto the 10 leading empirical orthogonal functions of ERA-Interim. Four significant clusters
235 are identified, corresponding to the usual four regimes, and each day of re-analysis and model data is then sorted into one of the
236 four regimes. From this, persistence probabilities can easily be computed.[¶] Visual inspection suggests that histograms of persistence
237 probabilities are distributed according to a reverse log-normal distribution (i.e. a distribution X such that $-X$ is log-normal). Figure 2
238 shows the result of fitting such a distribution to the estimates of persistence probabilities of the NAO+ regime.^{||} It can be clearly seen
239 that GloSea5 underestimates persistence, with too many short-lived NAO+ events. Similar behaviour is seen for the other regimes (not
240 pictured). We note that this was already observed independently in [Matsueda and Palmer \(2018\)](#), which shows examples of similar
241 errors in other models as well. While this observation suggests weak persistence is a notable model bias, we note that this could in
242 principle be attributable to any of the four suggested sources of error above, and therefore may be e.g. due to a weak signal rather than
243 weak regimes, as suggested in [Siegert *et al.* \(2016\)](#).

244 In the next section we will formulate a method to test the impact of this weak persistence on the Markov model, by systematically
245 relaxing the seasonal persistence probabilities of the GCM back towards 1/2. The idea is simple. For each DJF season to be simulated
246 in our hindcast, we randomly draw two persistence probabilities p_+ and p_- for the two regimes: these represent the true probabilities

[¶]This cluster analysis was carried out in collaboration with I. Mavilia (CNR); a detailed analysis of the full results will appear in a forthcoming paper.

^{||}This fitting was carried out using the Python package ‘scipy’, which has a standard routine for fitting log-normal distributions. By applying this to the transition probabilities, we obtain a reverse log-normal distribution of persistence probabilities.

247 that reality follow. We then generate a DJF mean for reality by taking a 90-day random walk generated with these probabilities, where
 248 the distributions of positive/negative daily NAO indices drawn from are fitted to re-analysis. To generate a single ensemble member
 249 prediction, we do the same thing, except the persistence probabilities are first relaxed back towards 1/2 using a scaling parameter k ,
 250 which we refer to as the ‘regime fidelity’ parameter. Setting $k = 0$ means the ensemble members will, on average, have persistence
 251 probabilities exactly 1/2, while $k = 1$ means the members will, on average, have persistence probabilities matching the true ones. In
 252 this way we can generate an ensemble prediction and simulate a hindcast, where the persistence probabilities are varying each year.
 253 For technical reasons it is easier to work with log-odds rather than probabilities, which makes the implementation of this idea slightly
 254 more involved. The full details are explained in the next section.

255 We have also tested a method aimed at simulating the impact of (3) and (4), which we did by systematically relaxing the persistence
 256 probabilities back towards a specified climatological mean of 0.8, thereby reducing interannual variability (whether as a result of errors
 257 in global dynamics or more complex internal dynamics). The results of the two experiments were qualitatively very similar, so we will
 258 restrict ourselves to the former. It is plausible that any bias which introduces errors in the persistence probabilities is likely to produce
 259 similar results.

260 3.3. Technical formulation

261 Simulating an ensemble forecast of the NAO in this framework will require the generation of randomly drawn persistence probabilities
 262 as well as the specification of the distributions X_+ and X_- . To facilitate the former, we always work with *log-odds* rather than
 263 probabilities when manipulating or drawing probabilities. From an assumed distribution of probabilities p (taking values between 0
 264 and 1), a new, more Gaussian distribution \hat{p} (taking values between $-\infty$ and $+\infty$) is obtained via the log-odds transformation

$$\hat{p} = \log\left(\frac{p}{1-p}\right).$$

265 Specific values \hat{p}_s drawn from this latter distribution correspond to the probability p_s from p given by

$$p_s = \frac{1}{1 + e^{-\hat{p}_s}}.$$

266 Figure 3 shows the histograms of the log-odds of persistence probabilities of NAO+ and NAO-, estimated with ERA-Interim. Note that
 267 here the NAO- regime is explicitly formed by combining, from the 4 regimes derived by K-means clustering, the 3 regimes that are not
 268 NAO+ as one. Using standard nomenclature, the true NAO- regime, along with the Atlantic Ridge and Blocking regimes, have all been
 269 combined into one regime, which we will refer to again as NAO-. The distributions appear to be fairly well approximated by normal
 270 distributions, with means (standard deviations) of 1.54 (0.54) and 2.30 (0.48) respectively. When transformed back to probabilities,
 271 these means correspond to 82% and 91% respectively.

272 In order to estimate the distributions X_+ and X_- we again use ERA-Interim. We simply look at the value of the NAO index** on
 273 each day for which the atmosphere was deemed to be in the NAO+ regime. The histogram of these numbers define the projection of
 274 the NAO+ regime onto the NAO index. Similarly we obtain the projection of the NAO- regime onto the NAO index. Figure 4 shows
 275 the result of this. As expected, the NAO+ projection is almost exclusively positive, while the NAO- projection is mostly negative, with
 276 some overlap onto the positive quadrant. These positive values are associated with events that really belong to either the Atlantic Ridge
 277 or Blocking regimes, which can be shown to have a more neutral projection onto the NAO index.

278 The asymmetry between the distributions, along with the fact that they are both notably skewed, takes us away from the idealized
 279 picture of section 3.1. As we are considering an idealized situation, we will artificially rectify this discrepancy. We will firstly ignore the
 280 skewness, and assume the NAO+ projection is normally distributed, and so set $X_+ \sim \mathcal{N}(1.0, 0.5)$. We then set X_- to be the symmetric
 281 image, $X_- \sim \mathcal{N}(-1.0, 0.5)$. We verified that the three core metrics we are interested in are insensitive to this simplification.

282 The simulation of an ensemble forecast of a single DJF season now proceeds as follows. We start by specifying two persistence
 283 probabilities p_{obs}^+ and p_{obs}^- for the NAO+ and NAO- regimes. These are to be interpreted as the ‘true’ probabilities for a given DJF
 284 season. These are obtained by drawing two random log-odds numbers L_{obs}^+ and L_{obs}^- from the fitted normal distributions of the logodds
 285 histograms from figure 3. These are then transformed back into actual probabilities as explained earlier.

286 We perform the same procedure to generate an ensemble forecast, with the only exception being the choice of persistence
 287 probabilities, which we will modify according to a parameter $0 \leq k \leq 1$, referred to as the *regime fidelity* factor, and a noise factor ϵ ,
 288 corresponding to variability between ensemble members. For every season, we now want to relax the models persistence probabilities,
 289 p_{mod}^+ and p_{mod}^- , back towards 1/2. Because $\log\left(\frac{1/2}{1-(1/2)}\right) = 0$ and the log-odds function is monotonic, this is equivalent to relaxing the
 290 log-odds of these two probabilities, L_{mod}^+ and L_{mod}^- , back towards 0. This is achieved by setting

$$L_{mod}^+ = k \cdot L_{obs}^+ + \epsilon^+ \quad (11)$$

$$L_{mod}^- = k \cdot L_{obs}^- + \epsilon^- \quad (12)$$

291 where ϵ^+ and ϵ^- are random draws from a normal distribution $\mathcal{N}(0, 0.5)$. This noise factor represents the variability between ensemble
 292 members, allowing for individual members to perform better or worse by chance. The standard deviation of 0.5 was chosen as it is
 293 approximately the standard deviation of the climatological distributions of ERA-Interim (see figure 3). That is, we assume that the
 294 distribution of persistence probabilities across the ensemble has a spread equal to that of climatology itself. In the bimodal system
 295 considered here, this amounts to a form of model error as well, since it is assumed that the correct persistence probabilities are fixed
 296 by the ambient forcing, and so in a perfect model, every member should have the same persistence probabilities, matching those of
 297 reality. However, this ‘perfection’ of the ensemble was considered unrealistic, and produces hindcast simulations where the ensemble

**Obtained from ERA-Interim as the leading EOF of geopotential height at 500 hPa over the Euro-Atlantic region.

298 has too small a variance. The climatological ensemble spread to persistence added here was considered a simple, pragmatic default
 299 assumption. The impact of this choice on the results will be discussed later. Once L_{mod}^+ and L_{mod}^- have been obtained in this manner,
 300 they are transformed back into actual probabilities to be fed into the Markov model.

301 Because, as mentioned above, the logistic function applied to 0 returns 1/2, a value of $k = 0$ corresponds to relaxing persistence
 302 probabilities back to 1/2 on average (with deviations from this mean given by the ϵ^+ and ϵ^- terms). When $k = 1$, the ensemble
 303 members have persistence probabilities on average equal to the true probabilities. Intermediate values scale linearly between these two
 304 extremes. This source of error ($k < 1$) can therefore be thought of as weak regime structure, with the potential wells of the two states
 305 being too shallow; persistence is low and therefore predictability is impaired.

306 Now, given a choice of the regime fidelity factor $0 \leq k \leq 1$ we can generate our ensemble prediction, where for a given season, each
 307 member is the result of running a 90-day Markov chain with parameters determined as above: **the persistence probabilities are therefore**
 308 **different in each season**. Mimicking the conditions in Dunstone *et al.* (2016), the hindcast length is set to 35 years and the ensemble
 309 size to 40.

310 Figure 5 verifies the theoretical predictions deduced in the previous section, showing seasonal mean pdfs computed with this Markov
 311 model. Plots (a) and (b) show a situation where $p_+ = 0.9$ and $p_- = 0.7$, so the preferred state is NAO+ to a relatively high level of
 312 predictability. In (a) there is no persistence bias ($k = 1$), while in (b) the regime fidelity is set to 0.5. One can see that the perfect skill
 313 pdf is very close to being strictly positive as expected, even when sampled over just 90 days. The insertion of error weakens this, with
 314 the pdf now still being mostly positive, but admitting a larger proportion of negative values as well. As a result, the ensemble mean is
 315 much smaller in magnitude; less than half as large as the perfect skill magnitude. Plots (c) and (d) show the same but for a year with
 316 very little predictability: $p_+ = 0.75$, $p_- = 0.7$, again with $k = 1$ on the left and $k = 0.5$ on the right. The pdfs are close to random noise
 317 with mean 0, with the $k = 1$ pdf having a mean of 0.09 and the $k = 0.5$ pdf having a mean of 0.04, again, less than half the magnitude
 318 of the perfect skill case.

319 3.4. Correlations and RPC: dependence on regime fidelity

320 Figure 6 shows the actual and model predictability for the Markov model as a function of regime fidelity k : shading gives an indication
 321 of sampling variability, given as two standard deviations from the mean values. For each value of regime fidelity factor k , 1000
 322 simulations of a 40-member ensemble hindcast were performed to generate these statistics. Horizontal lines indicate the DePreSys3
 323 values, i.e. actual predictability 0.62 and model predictability 0.18. Where these intersect the thick blue and red lines indicates where
 324 obtaining these values is expected. By intersecting the horizontal lines with the shading, we can isolate, for each value, a region
 325 of regime fidelity values for which obtaining either DePreSys3 value lies within the confidence interval. This allows us to isolate a
 326 critical region of skill values for which both DePreSys3 values can be plausibly obtained: this region is marked with vertical lines and
 327 corresponds approximately to $0.18 < k < 0.41$.

328 For example, when $k = 0.3$, the expected values are

$$329 \text{Corr}(EnsMean, Obs) = 0.63 (0.43, 0.83) \quad (13)$$

$$330 \text{Corr}(EnsMean, Mem) = 0.19 (0.11, 0.27) \quad (14)$$

$$331 RPC_{est} = 2.28 (1.46, 3.11)$$

332 where the bracketed values are the lower and upper bounds of the 95% confidence interval. The expected values are virtually the same
 333 as those obtained by DePreSys3.

334 Because the two correlations are not independent, and always come as a *pair* associated to a specific simulation, it is possible that
 335 while each individual number is easy to obtain, the pair itself is unlikely. To estimate the true joint probability, we have in figure 7
 336 explicitly plotted around 50000 such pairs, obtained from simulations where we restricted ourselves to values of k in the critical range
 337 only. It can be seen that the DePreSys3 pair lies at the centre of the cluster, indicating that obtaining this exact pair is very plausible
 338 with this model. In figure 8 we have plotted a typical example of a hindcast simulation with $k = 0.3$. Figure 9 shows the same, except
 339 we have normalized the ensemble mean and observations to have standard deviation 1. In this example, the three main metrics were
 340 0.55, 0.14 and 2.31. These figures should be compared to the equivalent figures 1a and 2a in Dunstone *et al.* (2016).

341 **Observe that in figure 6, actual predictability increased much more rapidly with k than model predictability does.** This can be
 342 understood conceptually in the following way. The seasonal mean is to a large extent captured by the preferred regime state, which is in
 343 turn determined, via equation (9), by the *difference* between the two persistence probabilities. The ensemble mean will, for a sufficiently
 344 large ensemble size, have persistence probabilities determined by equations (11) and (12). As such, both reality and the ensemble
 345 mean will, for a large ensemble, have the same preferred state in each season and correlate strongly with each other. However, many
 346 individual ensemble members will firstly, due to the internal variability, have sufficiently different persistence probabilities to ensure
 347 they have a different preferred regime. Secondly, other members, while still retaining the preferred regime of reality, will end up with
 348 the probabilities for both regime being closer together (both being close to 1/2): in this setting of reduced predictability such members
 349 may well end up realizing the wrong preferred regime state simply due to sampling variability. This would correspond, in figure 5(c),
 350 to a member ending up in the negative quadrant. The ensemble mean will correlate poorly with all such members. In total therefore,
 351 the ensemble mean will be expected to correlate more consistently strongly with reality than with individual ensemble members.

352 This is naturally reflected also in the *RPC*. Figure 10 shows the RPC_{est} as a function of regime fidelity k . With maximal regime
 353 fidelity ($k = 1$), the expected RPC_{est} is around 1.3. The fact that it isn't exactly 1 is due to the internal variability between ensemble
 354 members. If this were eliminated, so that all ensemble members have the exact same persistence probabilities, then the expected
 355 RPC_{est} with $k = 1$ is 1. Since around 75% of the discrepancy between the DePreSys3 value of 2.3 and the desired value of 1 is
 356 eliminated by improving the regime fidelity, we conclude that the dominant cause for obtaining $RPC_{est} > 1$ in this system is due to
 insufficient regime fidelity.

Observe that in figure 10, the change in slope at around $k = 0.2$ is due to the divergence of RPC_{est} from RPC_{true} . For $k < 0.2$, the
 model error has grown large enough to ensure that an ensemble mean of just 40 members no longer accurately represents reality. Figure

357 **11** shows RPC_{est} as a function of actual predictability. It can be seen that this divergence in the estimated RPC from true RPC begins
 358 to take effect for correlations of 0.4 and lower. This suggests that RPC estimates in situations where the ensemble mean correlation
 359 is less than 0.4 are unlikely to be meaningful; this is also reflected in the large errorbars in this region. Of particular importance here
 360 is that, as explained in section 2.2, the expected RPC_{est} at $k = 0$ here is around 0.85. Due to the large errorbars, it is therefore likely
 361 that, when estimating the RPC of a model with insufficient skill, one may compute $RPC_{est} \approx 1$ even when $RPC_{true} \gg 1$.

362 Our conclusion is that this idealized Markov model captures well the observed RPC behaviour of DePreSys3. Importantly, it does
 363 so even though the models internal level of noise, on *daily* timescales, is exactly identical to that of observations, both being measured
 364 by the standard deviation of X_+ . The apparently unrealistic amount of noise on *seasonal* timescales, as measured with the signal+noise
 365 model in Siegert *et al.* (2016), is in this context a purely emergent feature of model error. In other words, a weak signal with realistic
 366 noise on *daily* timescales can end up looking like a weak signal with a large amount of noise on *seasonal* timescales, when filtered
 367 through a regime system.

368 Finally, we note that a regime fidelity factor $k = 0.3$ corresponds roughly to the ensemble mean underestimating the true seasonal
 369 persistence probabilities (relative to ERA-Interim) by a factor of 0.8 on average. For comparison, the GloSea5 data considered in
 370 section 3.2 shows an underestimate by a factor closer to 0.9. While a fidelity factor $k = 0.41$, the highest possible factor still exhibiting
 371 the desired behaviour, brings this factor of 0.9 within the range of sampling variability, this is likely indicative of the two-state system
 372 being too crude an approximation.

373 3.5. Correlations and RPC : dependence on ensemble size

374 Figures 12 show how correlations grow as a function of ensemble size, where the regime fidelity factor k has been fixed at 0.3. The
 375 thick coloured lines show the expected actual and model predictability, while shading indicates the 95% confidence interval. Stipled
 376 lines show estimates of the same quantities as computed from raw DePreSys3 data. We see again that the Markov model captures the
 377 relationship very well.^{††}

378 In this model of the NAO, the benefit of increasing the ensemble size begins to saturate after around 80 members. With 100 members,
 379 the expected ensemble mean correlation is around 0.7 ± 0.2 . This is roughly the same as what can be obtained by setting the regime
 380 fidelity to 1, suggesting that one cannot expect to gain more skill by increasing the ensemble size further. By maximizing the regime
 381 fidelity, removing internal variability between members (so all members share the same persistence probabilities) and setting the
 382 ensemble size to 100, the expected correlation can be increased to around 0.75 ± 0.22 , the maximal level of skill obtainable here.
 383 Indeed, the expected correlation was found to be the same when increasing to 1000 ensemble members.

384 Figure 13 shows RPC_{est} versus ensemble size. As ensemble size increases, the ensemble mean becomes a better approximation
 385 of reality, and RPC_{est} becomes a better estimate of RPC_{true} . Since $RPC_{true} \geq RPC_{est}$, this explains why RPC_{est} increases with
 386 ensemble size. With 100 members, RPC_{est} is approximately 2.7. The flattening of the growth of RPC_{est} in figure 13 suggests that this
 387 is beginning to approach the true value RPC_{true} . With 1000 members, $RPC_{est} = 3.2$ with a 95% confidence interval of [2.48, 3.96],
 388 implying that the true RPC in this bimodal system is greater than 2.5 with high confidence.

389 A final point of interest here is that the scaling of these metrics with ensemble size provides one conceptually easy way to distinguish
 390 between different statistical models. For example, both the signal+noise model of Siegert *et al.* (2016), and our Markov model here,
 391 were fitted to a specific ensemble size, and therefore will be in good agreement with each other for this particular size. However, the
 392 manner in which they then scale when the ensemble size varies is different for the two models. In principle, this offers a method for
 393 distinguishing between the two models, though due to sampling uncertainty, the ensemble size would need to be on the order of 1000
 394 to robustly detect the difference.

395 4. Over or under-confident: RPC vs RMS error/spread

396 One way to interpret a model with $RPC > 1$ is that the model is ‘under-confident’, in the sense that individual members deviate from
 397 the ensemble mean more than is appropriate given the true level of predictability of the system. Another widely used method to assess
 398 if a model has an appropriate level of noise relative to its skill is to compare the average root-mean-square error of the ensemble mean
 399 to the total ensemble spread (Johnson and Bowler (2009)). A derivation of why these should be expected to match for a perfectly
 400 calibrated ensemble can be found e.g. in Fortin *et al.* (2014), which also includes precise definitions for the terms (in particular the
 401 correct definition of ensemble spread). If the error exceeds the spread, this could be understood as the model being ‘over-confident’
 402 (the ensemble pdf tends to be sharply centered at the wrong location), while if the spread exceeds the error, this is indicative of the
 403 model being ‘under-confident’ (the ensemble pdf is centered in roughly the correct place but is much wider than that of reality).

404 Figure 14 shows the RMS error over spread metric as a function of regime fidelity. It can be seen that while RPC_{est} is essentially
 405 always expected to be greater than 1, the error-spread metric changes notably as regime fidelity increases. For lower levels of regime
 406 fidelity, this metric is expected to be greater than 1, indicative of over-confidence, while the opposite is true for high levels of regime
 407 fidelity. In general therefore one cannot expect that RPC and error-spread metrics agree on whether a model is over or under-confident.
 408 By tweaking the parameters of the model one can easily obtain even more extreme examples. As expected, one can verify that when the
 409 model is statistically perfect ($k = 1$ and no variability between ensemble members), one finds that both RPC_{est} and RMS error/spread
 410 equal 1.

411 The scaling of error-spread with regime fidelity can be understood as follows. Due to equation (10), the variance of the ensemble is
 412 independent of the persistence probabilities, and is therefore going to be, on average, the same for every season. Therefore the decrease
 413 in error-spread is simply due to the ensemble mean error decreasing as regime fidelity increases.

414 If one computes these numbers for the DePreSys3 hindcasts, one finds that the ratio of root-mean-square error to ensemble spread
 415 is 0.87, indicative of a very small amount of model under-confidence. Since the overall ensemble spread matches that of observations

^{††}We remind the reader that the linear model in Siegert *et al.* (2016) is also able to capture this well, and so the Markov model’s good agreement here is not a unique feature.

well (see e.g. Dunstone *et al.* (2016)), this is due to the RMS error being slightly too small. With a regime fidelity of 0.3, the Markov model predicts this number to be around 1.2, suggesting that the Markov model has slightly too small a variance. For regime fidelity parameters closer to 0.4, the upper end of the critical range, one can however obtain 0.87 within the 95% confidence interval of the Markov model.

An important observation here^{‡‡} is that in situations where the magnitude of the ensemble mean is small, the RMS error will, by virtue of its definition, typically be close to the ensemble spread, and so the ratio will be close to 1. Therefore, depending on sample size, this metric may not easily detect any problems. On the other hand, the RPC_{est} metric, by virtue of the fact that correlations are insensitive to magnitude, will not have this short-coming. However, even this observation is subject to its own subtleties, visible by reference to figures 14 and 10. For example, when regime fidelity is 0.1, the RMS error divided by spread is robustly greater than 1, despite the fact that the signal in this situation is tiny. At the same time, the RPC_{est} metric cannot be clearly distinguished from 1, due to the poor limiting behaviour of the metric at very low levels of model skill. This again reinforces the point that these metrics, by measuring subtly different things will behave differently for imperfect models, making the attachment of meta-level descriptors (e.g. ‘over/under-confident’) tricky.

Finally, we note that it is the RMS error divided by spread metric which is most influenced by the added internal variability between persistence probabilities of individual ensemble members (the ϵ 's in equations (11) and (12)). Removing this variability entirely results in this metric being notably too large, due to a reduction in ensemble spread. This is due to the fact that because of equation (9), if all ensemble members have persistence probabilities with the correct relative magnitude, the ensemble members as well as the ensemble mean will tend to have the correct sign (unless the difference between the two probabilities is small). This is in contrast to what is observed in Dunstone *et al.* (2016), where ensemble members tend to inhabit both positive and negative values fairly evenly.

5. Conclusion

We constructed and analysed a statistical model aimed at capturing NAO predictability within an idealized bimodal regime setting. In this paradigm, predictability is provided by the models ability to faithfully represent the seasonally varying persistence probabilities between the different states, which in turn determine the preferred state in a given season. By fitting a single ‘regime fidelity’ parameter k , representing the GCMs ability to represent persistence adequately, we were able to use this model to accurately capture both the high level of skill and the large RPC_{est} values observed in the Dunstone *et al.* (2016). With this fitted parameter kept fixed, we showed that the scaling of these metrics with ensemble size also matches that of Dunstone *et al.* (2016) extremely closely. This suggests that the high skill of DePreSys3 together with its large RPC_{est} value could be because, while being able to detect changes in seasonal persistence probabilities well, it never the less retains systematic biases leading to a consistent under-estimation of persistence. Such model bias could, in this setting, be from a number of sources, including weak regime structure, poor propagation of external forcing/teleconnections, or generic errors in the initial conditions. As models are well known to have systematic problems with producing robust regimes compared to re-analysis, we speculate that alleviating this may be the most immediately accessible route towards even higher levels of NAO predictability and a better signal-to-noise ratio. While there is some evidence that both increased resolution and the implementation of stochastic physics schemes can alleviate such errors (see Dawson and Palmer (2015)), it is clear that there is still much work to be done in this area.

We emphasize to finish that the model tested is highly idealized, using a simplified bimodal setting. Many studies suggest there may be up to 4 regimes, where the model may be expected to have differing levels of skill and bias for all the associated 16 transition probabilities. In particular, small biases in multiple transitions may compound to produce large errors. A systematic study of this full regime system and the ability of models to capture this structure is a clear next step in illuminating this further, which the authors hope to address in future work. The results of this paper therefore should be understood primarily as a proof-of-concept result, demonstrating that the frequently observed model bias of weak regime persistence may be a crucial component to understanding NAO predictability and the ‘signal-to-noise’ paradox.

Acknowledgements

K.S. acknowledges funding from the European Commission under Grant Agreement 641727 of the Horizon 2020 research programme. T.N.P acknowledges funding by European Research Council grant number 291406. We are also particularly grateful to Nick Dunstone *et al.* who graciously offered the DePreSys3 hindcast data for our analysis. Conversations with David Stephenson and Fenwick Cooper were also of great value.

References

- Allan R, Ansell T. 2006. A new globally complete monthly historical gridded mean sea level pressure dataset (HadSLP2): 1850–2004. *Journal of Climate* **19**(22): 5816–5842, doi:10.1175/JCLI3937.1.
- Athanasiadis PJ, Bellucci A, Scaife AA, Hermanson L, Matera S, Sanna A, Borrelli A, MacLachlan C, Gualdi S. 2017. A multisystem view of wintertime NAO seasonal predictions. *Journal of Climate* **30**(4): 1461–1475, doi:10.1175/JCLI-D-16-0153.1.
- Corti S, Molteni F, Palmer TN. 1999. Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature* **398**(6730): 799–802, doi:10.1038/19745.
- Davini P, D’Andrea F. 2016. Northern Hemisphere atmospheric blocking representation in global climate models: Twenty years of improvements? *Journal of Climate* **29**(24): 8823–8840, doi:10.1175/JCLI-D-16-0242.1.
- Dawson A, Palmer TN. 2015. Simulating weather regimes: impact of model resolution and stochastic parameterization. *Climate Dynamics* **44**(7-8): 2177–2193, doi:10.1007/s00382-014-2238-x.
- Dawson A, Palmer TN, Corti S. 2012. Simulating regime structures in weather and climate prediction models. *Geophysical Research Letters* **39**(21), doi: 10.1029/2012GL053284.

^{‡‡}We thank an anonymous reviewer for pointing this out.

- 475 Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars AC, van de Berg
476 L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Hólm EV, Isaksen L, Kållberg P, Köhler M,
477 Matricardi M, McNally AP, Monge-Sanz BM, Morcrette JJ, Park BK, Peubey C, de Rosnay P, Tavolato C, Thépaut JN, Vitart F. 2011. The ERA-Interim
478 reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society* **137**(656): 553–597,
479 doi:10.1002/qj.828.
- 480 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
481 one year ahead. *Nature Geoscience* **9**(11): 809–814, doi:10.1038/ngeo2824.
- 482 Eade R, Smith D, Scaife A, Wallace E, Dunstone N, Hermanson L, Robinson N. 2014. Do seasonal-to-decadal climate predictions underestimate the
483 predictability of the real world? *Geophysical Research Letters* **41**(15): 5620–5628, doi:10.1002/2014GL061146.
- 484 Fortin V, Abaza M, Ancil F, Turcotte R. 2014. Why Should Ensemble Spread Match the RMSE of the Ensemble Mean? *Journal of Hydrometeorology* **15**(4):
485 1708–1713, doi:10.1175/JHM-D-14-0008.1, URL <http://journals.ametsoc.org/doi/abs/10.1175/JHM-D-14-0008.1>.
- 486 Greatbatch RJ. 2000. The North Atlantic Oscillation. *Stochastic Environmental Research and Risk Assessment* **14**(4): 0213–0242, doi:10.1007/s004770000047.
- 487 Johansson Å. 2007. Prediction skill of the NAO and PNA from daily to seasonal time scales. *Journal of Climate* **20**(10): 1957–1975, doi:10.1175/JCLI4072.1.
- 488 Johnson C, Bowler N. 2009. On the Reliability and Calibration of Ensemble Forecasts. *Monthly Weather Review* **137**(5): 1717–1720, doi:10.1175/
489 2009MWR2715.1, URL <http://journals.ametsoc.org/doi/abs/10.1175/2009MWR2715.1>.
- 490 Kumar A. 2009. Finite Samples and Uncertainty Estimates for Skill Measures for Seasonal Prediction. *Monthly Weather Review* **137**(8): 2622–2631, doi:
491 10.1175/2009MWR2814.1, URL <http://journals.ametsoc.org/doi/abs/10.1175/2009MWR2814.1>.
- 492 Kumar A, Peng P, Chen M. 2014. Is There a Relationship between Potential and Actual Skill? *Monthly Weather Review* **142**(6): 2220–2227, doi:
493 10.1175/MWR-D-13-00287.1, URL <http://journals.ametsoc.org/doi/abs/10.1175/MWR-D-13-00287.1>.
- 494 Marshall J, Kushnir Y, Battisti D, Chang P, Czaja A, Dickson R, Hurrell J, McCartney M, Saravanan R, Visbeck M. 2001. North Atlantic climate variability:
495 Phenomena, impacts and mechanisms. *International Journal of Climatology* **21**(15): 1863–1898, doi:10.1002/joc.693.
- 496 Matsueda M, Palmer T. 2018. Estimates of flow-dependent predictability of wintertime Euro-Atlantic weather regimes in medium-range forecasts. *Quarterly*
497 *Journal of the Royal Meteorological Society* doi:10.1002/qj.3265, URL <http://doi.wiley.com/10.1002/qj.3265>.
- 498 Müller WA, Appenzeller C, Schär C. 2005. Probabilistic seasonal prediction of the winter North Atlantic Oscillation and its impact on near surface temperature.
499 *Climate Dynamics* **24**(2-3): 213–226, doi:10.1007/s00382-004-0492-z.
- 500 Norris J. 1997. *Markov chains*. Cambridge University Press.
- 501 Palmer TN. 1999. A nonlinear dynamical perspective on climate prediction. *Journal of Climate* **12**(2): 575–591, doi:10.1002/j.1477-8696.1993.tb05802.x.
- 502 Raoult B, Bergeron C, Alós AL, Thépaut JN, Dee D. 2017. Climate service develops user-friendly data store. *ECMWF Newsletter* **151**, doi:10.21957/p3c285.
- 503 Scaife aa, Arribas a, Blockey E, Brookshaw a, Clark RT, Dunstone N, Eade R, Fereday D, Folland CK, Gordon M, Hermanson L, Knight JR, Lea DJ, MacLachlan
504 C, Maidens a, Martin M, Peterson aK, Smith D, Vellinga M, Wallace E, Waters J, Williams a. 2014. Skillful long range prediction of European and North
505 American winters. *Geophysical Research Letters* **5**: 2514–2519, doi:10.1002/2014GL059637. Received, URL <http://onlinelibrary.wiley.com/doi/10.1002/2014GL059637/full>.
- 506 Scaife AA, Woollings T, Knight J, Martin G, Hinton T. 2010. Atmospheric blocking and mean biases in climate models. *Journal of Climate* **23**(23): 6143–6152,
507 doi:10.1175/2010JCLI3728.1.
- 508 Shi W, Schaller N, Macleod D, Palmer TN, Weisheimer A. 2015. Impact of hindcast length on estimates of seasonal climate predictability. *Geophysical Research*
509 *Letters* **42**(5): 1554–1559, doi:10.1002/2014GL062829.
- 510 Siegert S, Stephenson DB, Sansom PG, Scaife AA, Eade R, Arribas A. 2016. A Bayesian framework for verification and recalibration of ensemble forecasts:
511 How uncertain is NAO predictability? *Journal of Climate* **29**(3): 995–1012, doi:10.1175/JCLI-D-15-0196.1.
- 512 Smith DM, Scaife AA, Eade R, Knight JR. 2016. Seasonal to decadal prediction of the winter North Atlantic Oscillation: Emerging capability and future
513 prospects. *Quarterly Journal of the Royal Meteorological Society* **142**(695): 611–617, doi:10.1002/qj.2479.
- 514 Straus DM, Corti S, Molteni F. 2007. Circulation regimes: Chaotic variability versus SST-forced predictability. *Journal of Climate* **20**(10): 2251–2272, doi:
515 10.1175/JCLI4070.1.
- 516 Woollings T, Hannachi A, Hoskins B, Turner A. 2010. A regime view of the North Atlantic oscillation and its response to anthropogenic forcing. *Journal of*
517 *Climate* **23**(6): 1291–1307, doi:10.1175/2009JCLI3087.1.
- 518 Woollings T, Hoskins B, Blackburn M, Berrisford P. 2008. A New Rossby WaveBreaking Interpretation of the North Atlantic Oscillation. *Journal*
519 *of the Atmospheric Sciences* **65**(2): 609–626, doi:10.1175/2007JAS2347.1, URL [http://journals.ametsoc.org/doi/abs/10.1175/](http://journals.ametsoc.org/doi/abs/10.1175/2007JAS2347.1)
520 [2007JAS2347.1](http://journals.ametsoc.org/doi/abs/10.1175/2007JAS2347.1).

522 6. Appendix A

We sketch how to derive the average proportion of time spent in each state of a bimodal Markov chain with persistence probabilities α and β , as given in section 3.1. Because Markov chains are usually discussed in the literature in terms of transition probabilities, we define $a = 1 - \alpha$ and $b = 1 - \beta$. Therefore we are dealing with a 2-state Markov chain with transition matrix

$$P = \begin{pmatrix} a & 1 - a \\ b & 1 - b \end{pmatrix}$$

Now recall, for example by combining theorem 1.7.7 and theorem 1.10.2 from Norris (1997), that in this setting, P has a unique invariant distribution, i.e. a vector (π_1, π_2) such that

$$(\pi_1 \quad \pi_2) P = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix}$$

523 where $\pi_1 + \pi_2 = 1$, and, furthermore, π_1 and π_2 are precisely the average proportion of time spent in the two states. This is a system
524 of linear equations which, due to the constraint $\pi_1 + \pi_2 = 1$, has a unique solution, which can readily be deduced with basic linear
525 algebra. Doing so and expressing the final answer in terms of persistence probabilities gives the expressions found in 3.1.

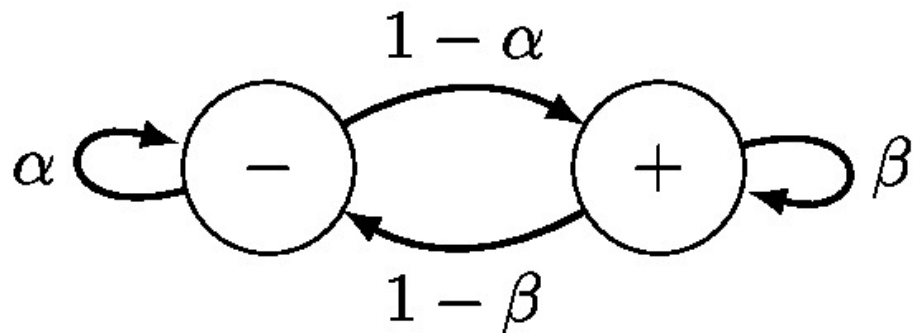


Figure 1. Illustration of a 2-state Markov chain with persistence probabilities α and β .

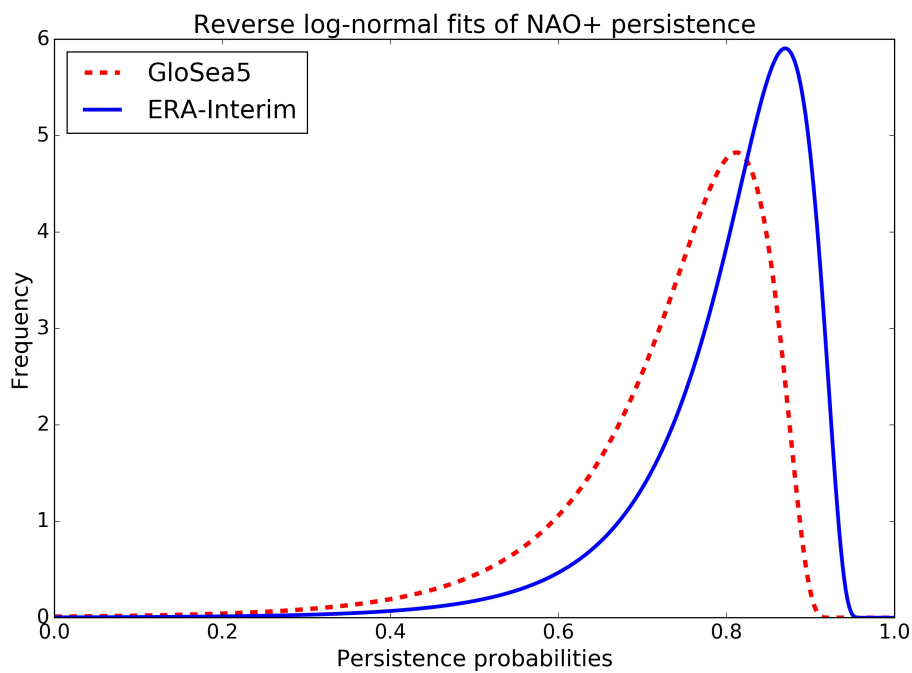


Figure 2. Reverse log-normal distributions fitted to estimates of NAO+ persistence probabilities for ERA-Interim (blue solid) and GloSea5 (red dotted).

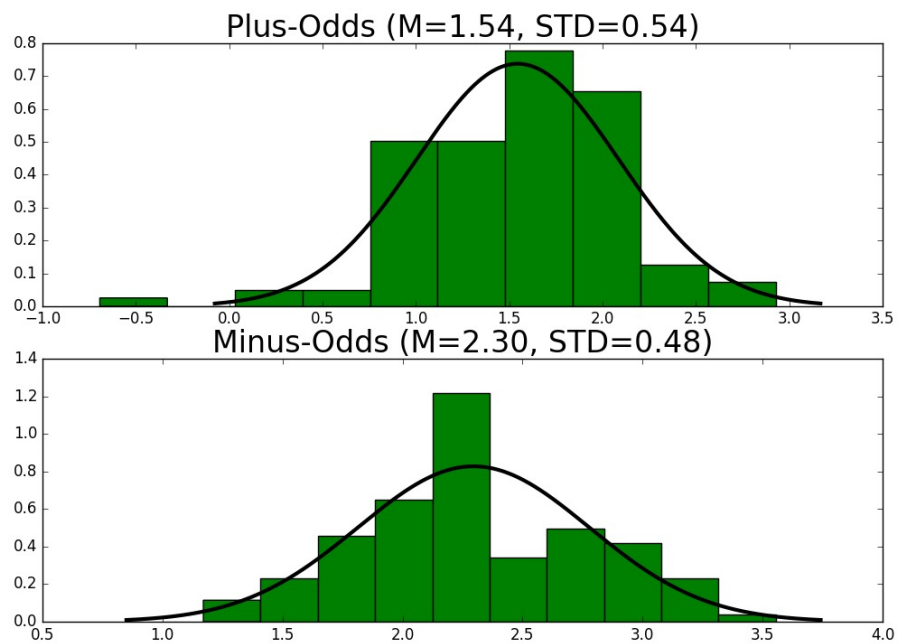


Figure 3. Histograms of log-odds of persistence probabilities of NAO+ (upper) and NAO- (lower), along with best-fit normal distributions (black curves) of both. Recall that NAO- here refers to the 3 standard regimes NAO-, Atlantic Ridge and Blocking all merged into one.

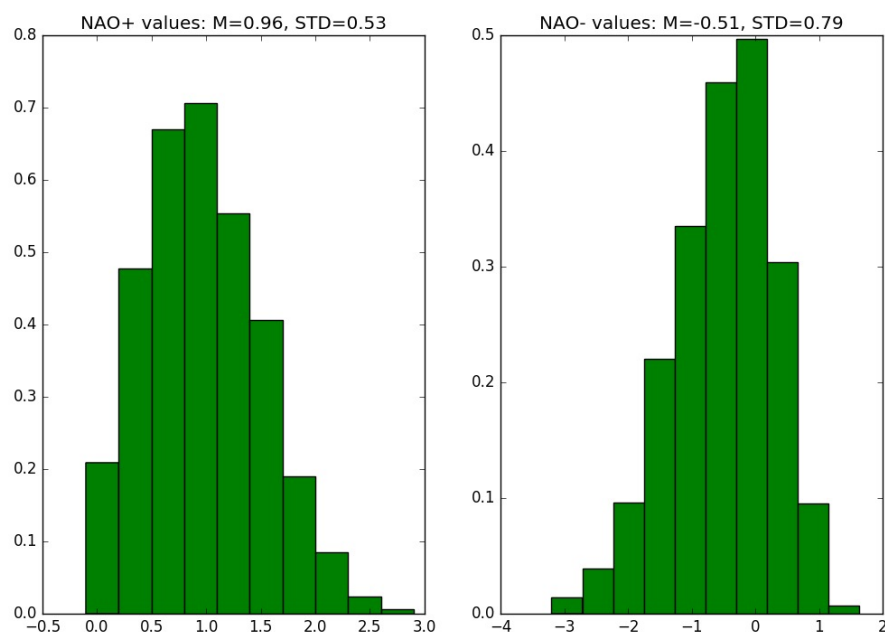


Figure 4. Projections of the NAO+ (left) and NAO- (right) regimes onto the NAO index, using ERA-Interim. Recall that NAO- here refers to the 3 standard regimes NAO-, Atlantic Ridge and Blocking all merged into one.

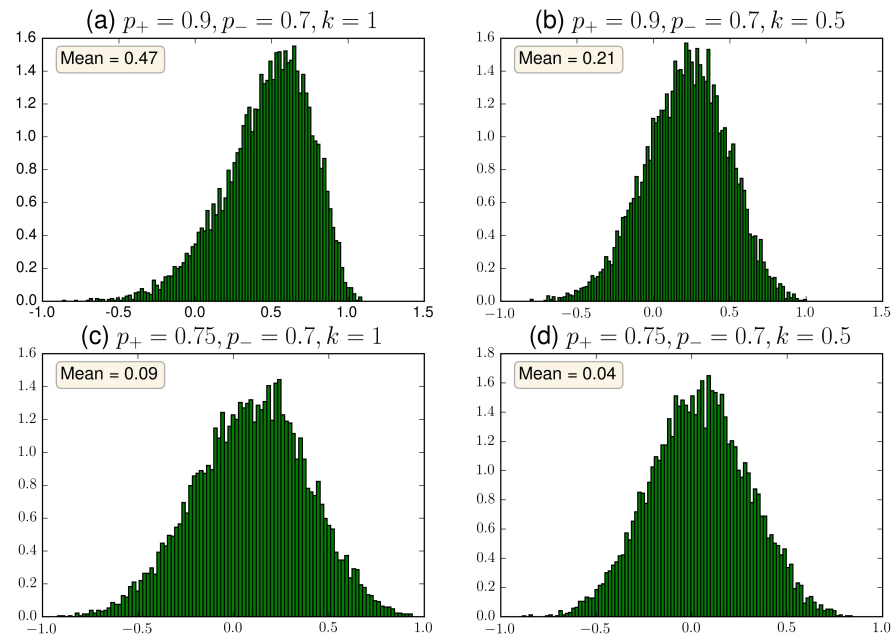


Figure 5. Histograms of seasonal mean ensembles using the Markov model. Plots (a) and (b) show a NAO+ year, with $p_+ = 0.9, p_- = 0.7$; in (a) regime fidelity is 1 (no persistence bias), in (b) the regime fidelity is 0.5. Plots (c) and (d) show a less predictable NAO+ year, with only a weak preference; $p_+ = 0.75, p_- = 0.7$. In (c) regime fidelity equals 1, in (d) regime fidelity equals 0.5. The histograms were produced from 10000 samples each.

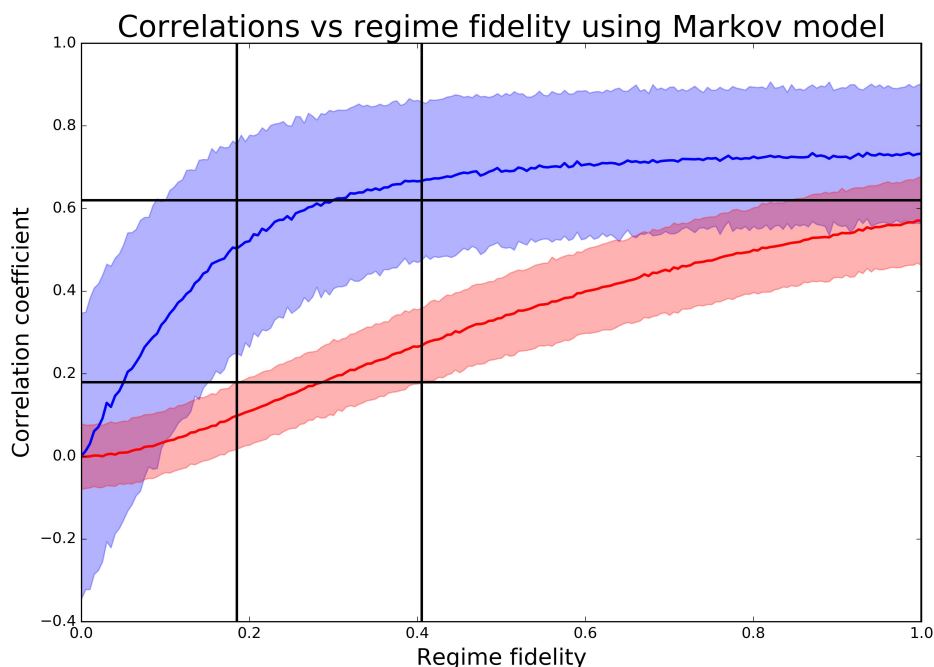


Figure 6. Correlations for the Markov model. $\text{Corr}(\text{EnsMean}, \text{Obs})$, the upper curve in blue, denotes the expected correlation between the ensemble mean and observations, while $\text{Corr}(\text{EnsMean}, \text{Mem})$, the lower curve in red, denotes the expected model predictability i.e. the average correlation between ensemble mean and individual members. Each correlation is plotted as a function of the regime fidelity parameter k . Shading indicates two standard deviations from the mean. Both the mean and standard deviations are computed, for each k , over a sample of size 1000. Horizontal lines indicate the values seen in DePreSys3 of 0.62 and 0.18, while vertical lines isolate a critical range for which these values are plausibly obtained (see main text).

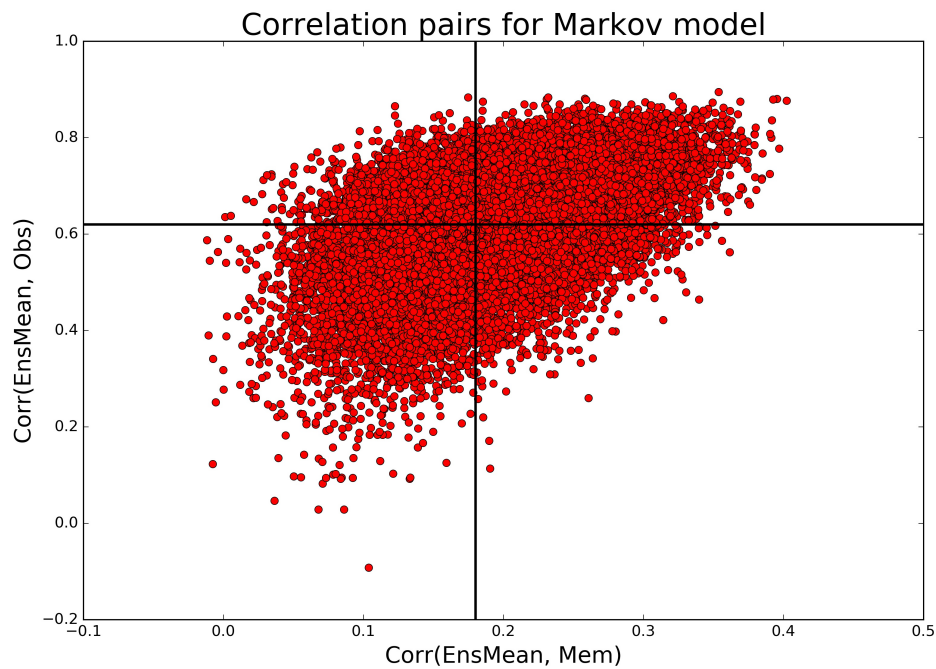


Figure 7. Actual pairs of correlations obtained from simulations using the Markov model. For a chosen value of regime fidelity k in the critical range (see main text), we ran 1000 simulations, and for each such simulation computed a single pair. Fifty choices of k were made, evenly sampled within the critical range, so the figure shows a total of 50000 such pairs. The intersection of the horizontal and vertical line marks the pair (0.18, 0.62) observed in DePreSys3.

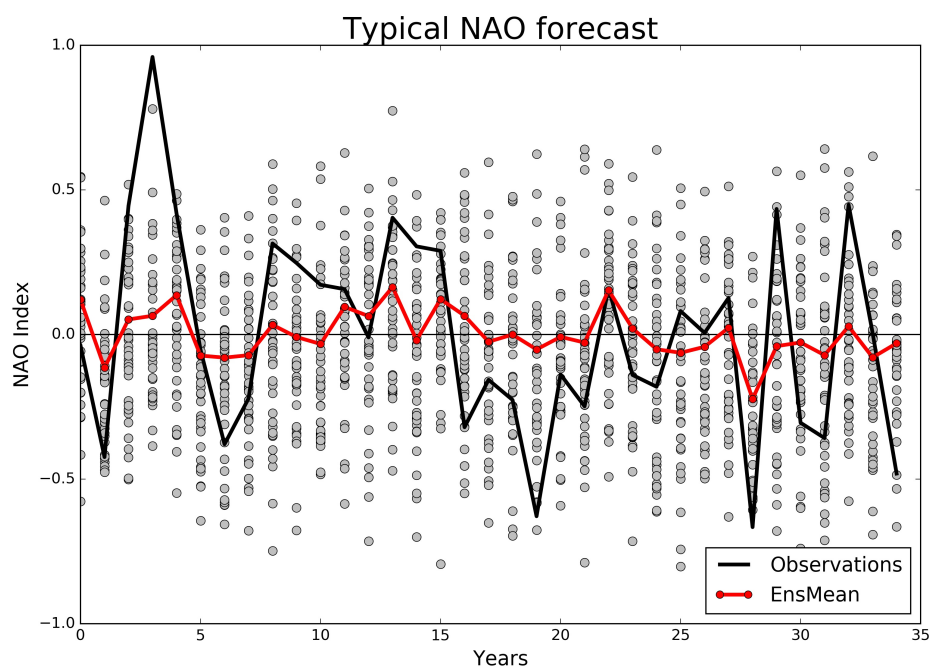


Figure 8. A typical example of a hindcast simulation using the Markov model with regime fidelity $k = 0.3$. Individual ensemble members are displayed as grey dots, with the ensemble mean in red (dotted). In black is the randomly generated ‘observations’ for this simulation. In this hindcast, actual predictability was 0.55, model predictability was 0.14, and RPC_{est} was 2.31.

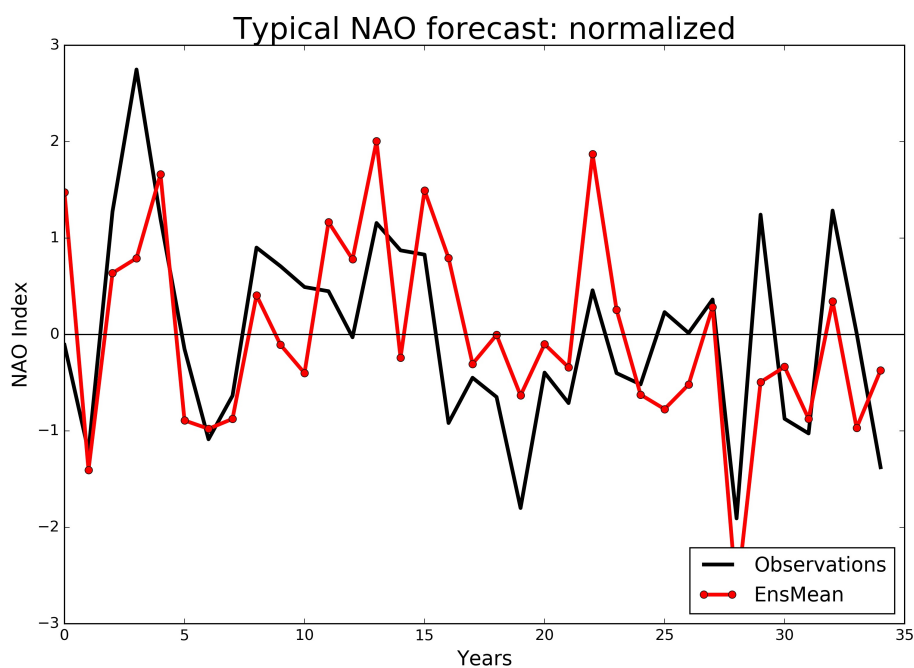


Figure 9. The same forecast as in figure 8, but now ensemble mean (red dotted) and reality (black) have been normalized to have standard deviation 1.

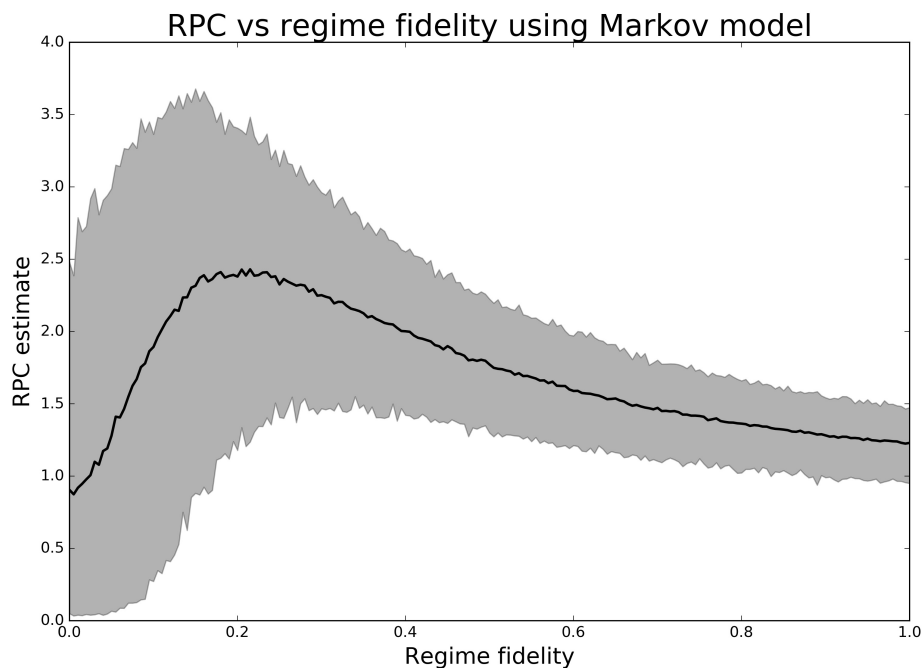


Figure 10. RPC_{est} as a function of regime fidelity k (see main text) using the Markov model. The thick line shows the expected value for the given choice of skill, while shading denotes the 95% confidence interval. For each skill value, 1000 simulations were used to generate these statistics.

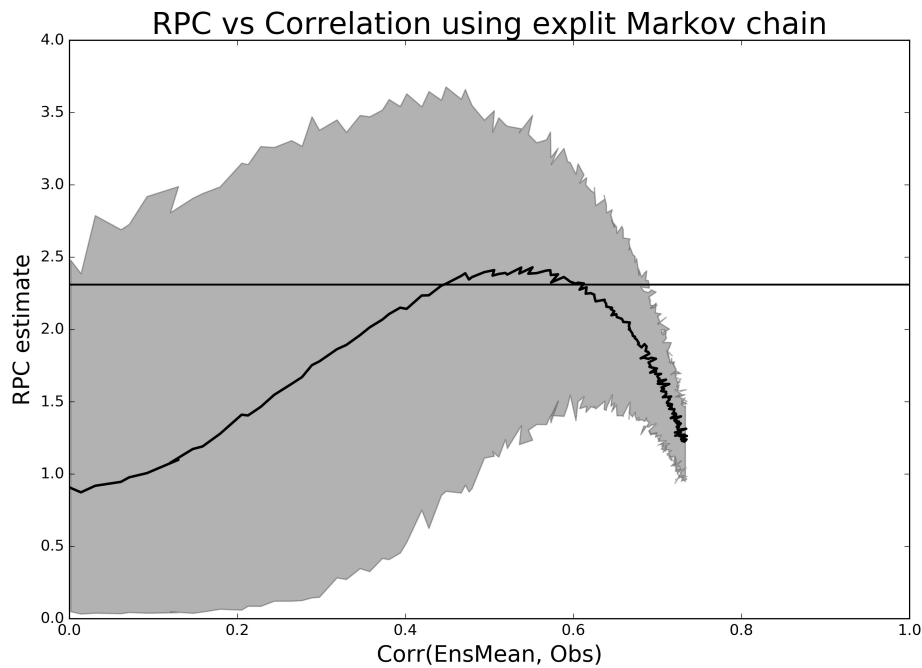


Figure 11. RPC_{est} as a function of ensemble mean correlation, using the Markov model. The thick line shows the expected value for the given level of correlation, while shading denotes the 95% confidence interval. For each correlation coefficient, 1000 simulations were used to generate these statistics. The horizontal line marks the value of 2.31 obtained by DePreSys3.

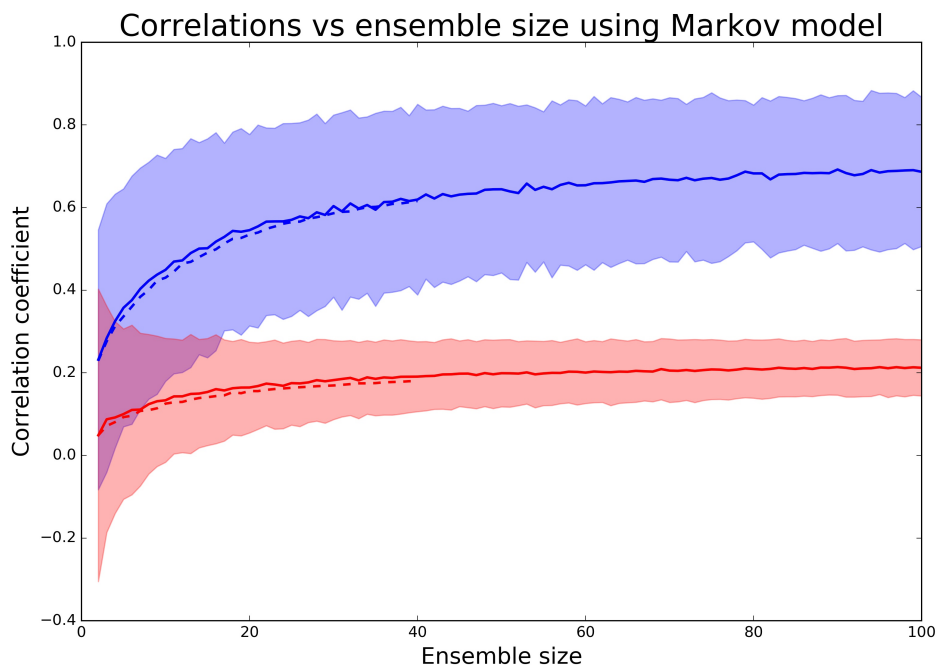


Figure 12. Correlations as a function of ensemble size. Actual predictability, the upper curve in blue, shows the expected correlation between the ensemble mean and observations, while model predictability, the lower curve in red, shows the average correlation between ensemble mean and individual members. Thick lines show the expected value for the given ensemble size, while shading denotes the 95% confidence interval; for each choice of ensemble size, 1000 simulations were performed to generate the statistics. Regime fidelity was fixed at $k = 0.3$. The stippled lines show estimates of the same quantities from DePreSys3.

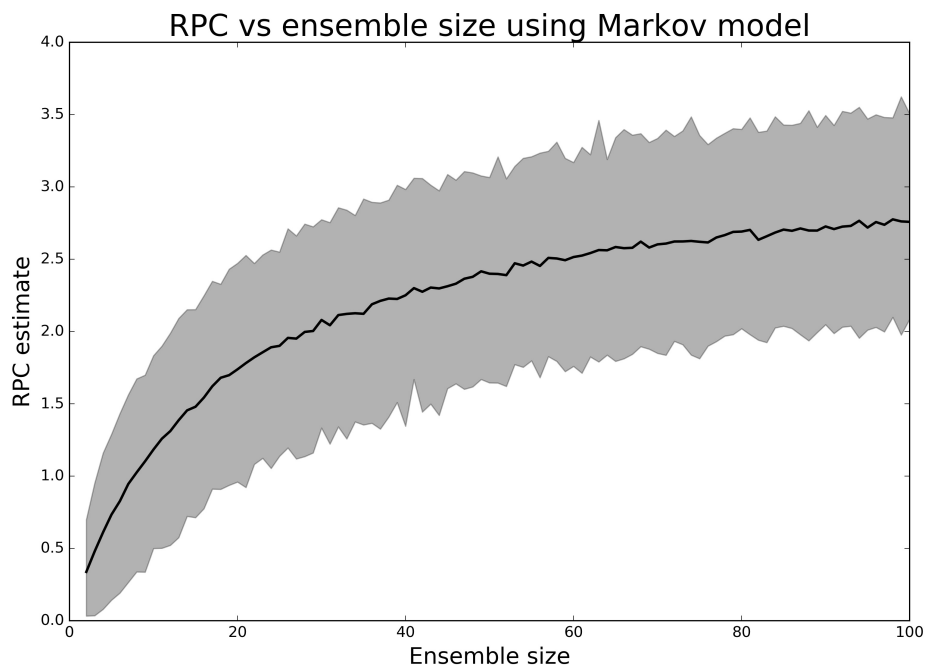


Figure 13. RPC_{est} as a function of ensemble size, using the Markov model. The thick line shows the expected value for the given choice of skill, while shading denotes the 95% confidence interval. For each skill value, 1000 simulations were used to generate these statistics. The regime fidelity was fixed at $k = 0.3$.

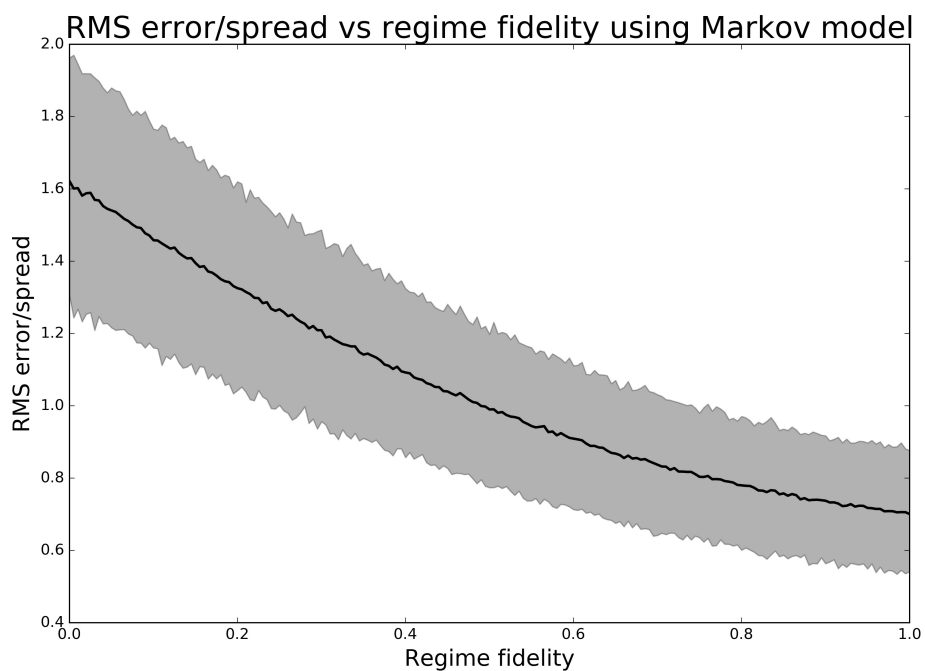


Figure 14. Root-mean-square ensemble mean error divided by ensemble spread for the Markov model, for different values of regime fidelity k . Shading indicates the 95% confidence interval and the solid line the mean. For each k , 1000 simulations were run to obtain these statistics.