A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability?

Stefan Siegert[1]*, David B. Stephenson[1], Philip G. Sansom[1],
Adam A. Scaife[2], Rosie Eade[2], Alberto Arribas[2]

[1] *University of Exeter, Exeter, United Kingdom*
[2] *Met Office, Hadley Centre, Exeter, United Kingdom*

September 24, 2015

Abstract

Predictability estimates of ensemble prediction systems are uncertain due to limited numbers of past forecasts and observations. To account for such uncertainty, this paper proposes a Bayesian inferential framework that provides a simple 6-parameter representation of ensemble forecasting systems and the corresponding observations. The framework is probabilistic, and thus allows for quantifying uncertainty in predictability measures such as correlation skill and signal-to-noise ratios. It also provides a natural way to produce recalibrated probabilistic predictions from uncalibrated ensembles forecasts. The framework is used to address important questions concerning the skill of winter hindcasts of the North Atlantic Oscillation for 1992-2011 issued by the Met Office GloSea5 climate prediction system. Although there is much uncertainty in the correlation between ensemble mean and observations, there is strong evidence of skill: the 95% credible interval of the correlation coefficient of $[0.19, 0.68]$ does not overlap zero. There is also strong evidence that the forecasts are not exchangeable with the observations: With over 99% certainty, the signal-to-noise ratio of the forecasts is smaller than the signal-to-noise ratio of the observations, which suggests that raw forecasts should not be taken as representative scenarios of the observations. Forecast recalibration is thus required, which can be coherently addressed within the proposed framework.

1 Introduction

Recent studies [Riddle et al., 2013, Scaife et al., 2014, Kang et al., 2014] corroborate that state-of-the-art atmosphere-ocean models produce skillful predictions

---

*Corresponding author: Stefan Siegert, Exeter Climate Systems, University of Exeter, North Park Road, Exeter, EX4 4QE, United Kingdom, Email: s.siegert@exeter.ac.uk

of climate variability on seasonal time scales. The performance of such forecasting systems is generally estimated by calculating summary sample statistics, such as correlation, over a limited sample of past forecasts and corresponding observations [e.g., Goddard et al., 2013]. It is then assumed that future forecasts will exhibit similar performance characteristics [Otto et al., 2012].

However, such measures-oriented forecast verification [Jolliffe and Stephenson, 2012] provides no inherent information about uncertainty in the reliability and skill of the forecast. Uncertainty in forecast quality estimates can be substantial for the small time samples and ensemble sizes typical of climate prediction systems. Without proper uncertainty quantification, it is difficult to address important questions for the development and use of climate services such as:

1. Could the observed skill be due to chance sampling, i.e., natural variability in the observed events and ensemble of forecasts?

2. How might the skill vary for a different non-overlapping time period (e.g., in the future)?

3. How might the skill vary if a new set of ensemble forecasts were generated over the same hindcast period?

4. Are the forecasts exchangeable with the observations, i.e., do the individual model forecasts have similar properties to the observations?

5. How can non-exchangeable ensemble forecasts be used to create a reliable probability forecast of future observations?

To address such questions it is helpful to propose a statistical model capable of representing the joint distribution of R members of an ensemble forecast, $x_{t,1}, \ldots, x_{t,R}$, and their verifying observation y_t over a set of times $t = 1, \ldots, N$.

The importance of an explicit statistical model has been recognized for climate change projections, where statistical models have been used to formalize assumptions about climate model output and the observed present and future climate [Tebaldi et al., 2005, Sansom et al., 2013]. Chandler [2013] argues that a statistical model can make all subjective model assumptions (and limitations) explicit, which leads to transparency in subsequent analyses. The importance of statistical modeling has also been recognized for weather and seasonal climate forecasting, where the prevailing application is to specify the forecast distribution, i.e., the conditional distribution of the observations, given the raw numerical model output. Statistical modeling in this context is referred to as forecast recalibration; the goal is to eliminate systematic biases from the numerical model output to improve forecast accuracy. Commonly used methods for forecast recalibration include Model Output Statistics [MOS, Glahn and Lowry, 1972], ensemble dressing [Wang and Bishop, 2005], and Non-homogeneous Gaussian Regression [NGR, Gneiting et al., 2005]. In these recalibration frameworks, the forecasts are not perceived as random quantities, and the full joint distribution of forecasts and observations is not specified. The present study highlights the

benefits of modelling the full joint distribution of forecasts and observations, rather than only the conditional forecast distribution. The joint distribution captures the variability and dependencies of numerical model forecasts and verifying observations, and thus contains useful information for forecast verification. The approach of evaluating forecast quality from the joint distribution is known as distributions-oriented verification [Murphy and Winkler, 1987]. It has not been widely applied because sample sizes of hindcast data sets are usually too small to estimate the joint distribution in sufficient detail. Parametric modeling has been identified as a useful approach to overcome the curse of dimensionality for distributions-oriented forecast verification [e.g., Murphy and Wilks, 1998, Bradley et al., 2004]. In this study we specify the joint distribution of forecasts and observations using a parametric statistical model. The parameters have to be estimated from a small data set of past forecasts and observations, and are therefore uncertain. We therefore advocate a framework which uses Bayesian inference to simultaneously estimate the parameters and quantify their uncertainty. We show how a Bayesian framework can be applied to verification and recalibration of ensemble forecasts based on a small hindcast data set.

So how should one model an ensemble forecasting system so as to capture the relevant dependencies and variations in forecasts and observations? In this paper, we study a signal-plus-noise model for an ensemble of runs from a numerical forecast model and the corresponding observations. The statistical model assumes the existence of a predictable "signal" which generates correlation between forecast model runs and observations, as well as the existence of unpredictable "noise", which leads to internal variability and random forecast errors. "Signal" and "noise" are modeled as independent Normally distributed random variables. The members of the numerical forecast ensemble are assumed to be exchangeable with one another, i.e. statistically indistinguishable, but not necessarily exchangeable with the observations. Possible violations of exchangeability captured by the chosen signal-plus-noise model include a constant bias of the mean, a linear transformation of the predictable signal, as well as differing signal-to-noise ratios. The signal-plus-noise model is related to the statistical models used by Murphy [1990], Kharin and Zwiers [2003], Weigel et al. [2009], and Kumar et al. [2014]. We discuss these in more detail in sec. 2, describe new methods for estimating the model parameters, and present novel applications of the signal-plus-noise model to verification and recalibration of seasonal climate forecasts.

In sec. 3, the proposed statistical framework is used to analyze recent North Atlantic Oscillation (NAO) hindcasts made with the Met Office GloSea5 seasonal climate prediction system [MacLachlan et al., 2014, Scaife et al., 2014]. We demonstrate how the framework allows us to coherently address questions 1-5 above, i.e., to analyze uncertainty in correlation skill, assess the exchangeability of forecasts and observations, and transform raw ensemble forecasts into recalibrated predictive distribution functions.

3

2 A signal-plus-noise model for ensemble forecasts

The statistical model used here is motivated by a simple interpretation of ensemble forecasts in the climate sciences, which assumes that observations and forecasts share a common predictable component (the signal), and unpredictable discrepancies arise due to model errors, internal variability, measurement error etc. (the noise). Although the same or similar statistical models have been used in previous studies (summarized in sec. 2.2), we will provide a detailed discussion of the underlying statistical assumptions and their implications.

2.1 The signal-plus-noise model

Let y_t be the observation at time t, and $x_{t,r}$ the ensemble member ("run") r at time t. The time t assumes values $1, \cdots, N$, and the ensemble run index r assumes values $1, \cdots, R$. The model equations are:

$$y_t = \mu_y + \quad s_t + \epsilon_t \tag{1a}$$

$$x_{t,r} = \mu_x + \beta s_t + \eta_{t,r} \tag{1b}$$

where μ_y, μ_x and β are constants, and s_t, ϵ_t, and $\eta_{t,r}$ are assumed to be independent Normal random variables with mean zero and constant variances σ_s^2, σ_ϵ^2, and σ_η^2, respectively.

The marginal expected values of the observations y_t and the ensemble members $x_{t,r}$ are equal to μ_y and μ_x, respectively. The random variable s_t describes an unobservable "predictable signal" shared between forecasts and observations. The coupling parameter β determines the sensitivity of the forecasts to the predictable signal. The random variable ϵ_t models the unpredictable component of observed climate, or "weather noise", and the random variable $\eta_{t,r}$ models ensemble variability, or "model noise".

The model Eq. (1) includes a number of assumptions about the forecasts and observations. The data are Normally distributed, and forecasts and observations at different times are conditionally independent, given the model parameters. Forecasts and observations share a common source of variability, which is modeled by the random variable s_t. The ensemble members are statistically exchangeable with one another, but are generally not exchangeable with the observation. There exist systematic and/or random discrepancies between model runs and observations, which includes the possibility of a constant model bias ($\mu_x - \mu_y \neq 0$), and possibly different strengths of the predictable signal and unpredictable noise in forecast and observation ($\beta \neq 1$, $\sigma_\epsilon \neq \sigma_\eta$).

We have argued in the Introduction that it is useful to specify a model for the full joint distribution of forecasts and observations. Under the model given by Eq. (1), forecasts and observations are distributed as a multivariate Normal distribution:

$$(y, x_1, \cdots, x_R)^T \sim \mathcal{N}(\mu, \Sigma) \tag{2}$$

4

with $(R+1)$-dimensional mean vector

$$\mu = (\mu_y, \mu_x, \cdots, \mu_x)^T. \tag{3}$$

The $(R+1) \times (R+1)$ dimensional covariance matrix Σ has entries

$$var(y) = \sigma_s^2 + \sigma_\epsilon^2, \tag{4a}$$
$$var(x_i) = \beta^2 \sigma_s^2 + \sigma_\eta^2, \tag{4b}$$
$$cov(x_i, x_j) = \beta^2 \sigma_s^2 \ (i \neq j), \text{ and} \tag{4c}$$
$$cov(x_i, y) = \beta \sigma_s^2. \tag{4d}$$

for all $i, j = 1, \ldots, R$. Therefore, the model Eq. (1) can be considered as a simplified parametrization of a covariance matrix of jointly Normal ensemble members and observations, which assumes exchangeability among the ensemble members. By modeling the $R+1$ observable random variables y_t and $x_{t,r}$ by an unobservable latent variable s_t, the number of free parameters in the covariance matrix Σ is reduced from $(R+1)(R+2)/2$ to only 4. Invoking a latent variable provides a parsimonious description of the joint distribution of forecasts and observations. Note further that the variance of the ensemble mean is given by

$$var(\bar{x}) = \beta^2 \sigma_s^2 + \frac{1}{R}\sigma_\eta^2, \tag{5}$$

and that the covariance $cov(x_i, y)$ between observations and individual ensemble members is equal to the covariance $cov(\bar{x}, y)$ between observations and the ensemble mean. The correlation skill of the ensemble mean can thus be expressed in terms of the model parameters by

$$\rho = \frac{cov(\bar{x}, y)}{\sqrt{var(\bar{x})var(y)}} = \frac{\beta \sigma_s^2}{\sqrt{(\beta^2 \sigma_s^2 + \sigma_\eta^2/R)(\sigma_s^2 + \sigma_\epsilon^2)}}. \tag{6}$$

The model parameters can be used to assess further aspects of the quality of the forecasting system. The forecasts are exchangeable with the observations if and only if $\mu_x = \mu_y$, $\beta = 1$, and $\sigma_\epsilon = \sigma_\eta$. If these conditions are met, the ensemble forecast is perfectly reliable, i.e., the observation is indistinguishable from the ensemble members, and the individual ensemble members can be taken as representative scenarios for the observation. If the forecast is reliable in the above parametric sense, the additional criterion $\sigma_\epsilon = \sigma_\eta = 0$ indicates a perfect deterministic forecast; all ensemble members are then always exactly equal to the observation. If, on the other hand, either $\beta = 0$ or $\sigma_s = 0$, there is no systematic relation between the forecasts and observations, i.e., the forecasts have no skill. The forecasts are marginally calibrated, i.e., forecast and observed climatology are equal, if $\mu_x = \mu_y$ and $\beta^2 \sigma_s^2 + \sigma_\eta^2 = \sigma_s^2 + \sigma_\epsilon^2$.

The variable s_t, referred to as "the predictable signal", requires careful interpretation. Essentially, this latent variable is a model construct that provides covariance - it cannot be directly observed. However, for climate predictions

the concepts of "signal" and "noise" can be (and have been) given a physical interpretation (e.g., Madden [1976], Von Storch and Zwiers [2001, sec. 17.2.2] or Eade et al. [2014]). The predictable signal can be understood as the slowly varying component of weather related to longer time scale processes (e.g., ocean circulation). The noise is interpreted as weather variability, which cannot be predicted deterministically on time scales of more than a few days. It should be noted that the signal estimated here is a property of the observations and the forecasts, and it is not a unique property of the "real world". Different forecasting models for the same observation can give rise to different "signals".

2.2 Related statistical models

Related models have been widely used for statistical data analysis, for example, in structural equation modeling [Pearl, 2000], factor analysis [Everitt, 1984], latent variable modeling [Bartholomew et al., 2011], and measurement error models, also known as error-in-variables models [Fuller, 1987, Buonaccorsi, 2010]. The same or similar models as our signal-plus-noise model Eq. (1) have also been used to investigate seasonal to decadal climate predictability. Kharin and Zwiers [2003] apply the signal-plus-noise model to seasonal climate forecast variability. Like the present study, Kharin and Zwiers [2003] use the model to study the relationship between variability and predictability, and also use the explicit statistical assumptions to calibrate imperfect ensemble forecasts to improve probabilistic forecast skill. Their parameter estimation is essentially based on the method of moments, and parameter uncertainty is not quantified. The present study extends Kharin and Zwiers [2003] by carefully quantifying uncertainty in the statistical model parameters as well as all derived quantities, and by incorporating this uncertainty in distributions-oriented forecast verification and forecast recalibration. More recently, Kumar et al. [2014] used the signal-plus-noise model to study the relationship between perfect skill and actual skill in seasonal ensemble forecasts. They show that perfect skill, i.e., the ability of the ensemble to predict its own realizations can be lower than actual skill, i.e., the ability of the ensemble to predict the real system. We will address actual and perfect model predictability in sec. 3.5, where we study signal-to-noise ratios in forecasts and observations. Unlike Kumar et al. [2014], the present study quantifies uncertainty in the signal-to-noise ratios.

The proposed signal-plus-noise model also relates to previous frameworks used to interpret ensembles of climate projections (see Stephenson et al. [2012] and references therein). Rougier et al. [2013] apply a latent variable models to infer future climate from a collection of exchangeable climate model runs. Chandler [2013] provides a statistical framework for multi-model ensembles, where runs from one climate model are non-exchangeable with runs from different climate models, and non-exchangeable with the observations. A related Bayesian framework is used by Tebaldi et al. [2005], who assume different values of model parameters for present and future climate. Annan and Hargreaves [2010] work under the assumption that ensemble forecasts and observations are fully statistically exchangeable; their model is thus a special case of the signal-plus-noise

6

model with $\beta = 1$, $\mu_x = \mu_y$, and $\sigma_\epsilon = \sigma_\eta$.

A noteworthy modification was studied by Weigel et al. [2009]. The observation is similarly decomposed into signal plus noise, but the ensemble members are modeled by adding a common random error term d_t as well as individual error terms $\eta_{t,r}$ to the predictable signal variable:

$$y_t = s_t + \epsilon_t \tag{7a}$$

$$x_{t,r} = s_t + d_t + \eta_{t,r}. \tag{7b}$$

We note that this additive model implies that the covariance between ensemble members is $cov(x_i, x_j) = \sigma_s^2 + \sigma_d^2$, and that the covariance between ensemble members and observations is $cov(x_i, y) = \sigma_s^2$, which implies that $cov(x_i, y)$ can never be negative, and $cov(x_i, x_j)$ can never be smaller than $cov(x_i, y)$. Both scenarios are, however, conceivable in real systems and should at least be allowed by a statistical model. Eq. (4) shows that model Eq. (1) does not impose these two restrictions; the only similar restriction is that, according to Eq. (4c), $cov(x_i, x_j)$ is always positive.

2.3 Parameter estimation

It is possible to calculate point estimates of the model parameters using the method of moments. This makes use of the first and second sample moments of the data and equates them with the corresponding expected values in Eq. (4). The estimating equations are given in appendix C. Such moment estimators are discussed by Moran [1971] (in the context of linear structural relationships) who notes that, if σ_η^2 were known exactly, then the moment estimators are also the maximum-likelihood estimators, and that complications can arise due to negative variance estimates which require modifications of the estimator equations. Point estimates obtained by method of moments or maximum likelihood estimation are prone to sampling uncertainty, especially for the small sample sizes typical of climate prediction systems. It is therefore important to quantify uncertainty in the model parameters, using either resampling methods such as the bootstrap [Efron and Tibshirani, 1994], by frequentist variance estimators or confidence intervals [e.g., Fuller, 1987], or by Bayesian estimation which we use here.

In Bayesian statistics degrees of certainty and uncertainty are expressed by conditional probabilities, and probabilities are manipulated based on the principle of coherence, i.e. by using only the addition and multiplication rule of probability calculus [Jaynes, 2003, Gelman et al., 2004, Lindley, 2006, Robert, 2007]. For the present study, the main object of interest for Bayesian inference is therefore the joint conditional probability distribution over all unknown quantities (i.e., the model parameters), conditional on all known quantities (i.e., the hindcast data and observations). From this posterior distribution we can derive point estimators, for example, the posterior mean or mode, and uncertainty intervals, for example the 95% parameter values with highest posterior density. Denote by $\theta = \{\mu_x, \mu_y, \beta, \sigma_s, \sigma_\epsilon, \sigma_\eta\}$ the collection of unknown parameters of

the signal-plus-noise model, by $s = \{s_1, \cdots, s_N\}$ the unknown values of the latent signal variable, and by $\{x, y\} = \{x_{t,1}, \cdots, x_{t,R}, y_t\}_{t=1}^N$ the collection of known forecasts and observations from a hindcast experiment. The desired posterior distribution for Bayesian estimation is thus $p(\theta, s | x, y)$. Derivation of the posterior distribution requires the specification of a prior probability distribution $\pi(\theta, s)$ over the unknown quantities, which factors into $\pi(\theta) \prod_{t=1}^N \pi(s_t | \sigma_s)$ in our model. The prior distribution can be used to incorporate a-priori information about the modeled data into the inference process (we discuss the prior distribution for our analysis in sec. 3.2). Furthermore, the likelihood function is required, which is the probability of the data, given specified values of the model parameters. The likelihood function, denoted by $\ell(x, y | \theta, s)$ can be calculated from Eq. (1) using the distribution law of the Normal distribution, and the independence assumption:

$$
\ell(x, y | \theta, s) = \prod_{t=1}^N \left\{ p(y_t | \theta, s_t) \prod_{r=1}^R p(x_{t,r} | \theta, s_t) \right\}
$$
$$
= \left(2\pi\sigma_\epsilon^2\right)^{-N/2} \left(2\pi\sigma_\eta^2\right)^{-NR/2} \exp\left\{ -\frac{1}{2} \sum_{t=1}^N \left[\left(\frac{y_t - (\mu_y + s_t)}{\sigma_\epsilon}\right)^2 + \sum_{r=1}^R \left(\frac{x_{t,r} - (\mu_x + \beta s_t)}{\sigma_\eta}\right)^2 \right] \right\}.
$$
$$
(8)
$$

Using the likelihood function and the prior distribution, the posterior distribution is then formally calculated by Bayes rule

$$
p(\theta, s | x, y) \propto \ell(x, y | \theta, s) \pi(\theta, s) \qquad (9)
$$

where the proportionality constant is independent of θ and s, and depends only on the data.

A closed form expression for the joint posterior distribution using arbitrary prior distributions is not available. For this paper, we have thus approximated a fully Bayesian analysis by Markov-chain Monte-Carlo (MCMC) integration [Brooks et al., 2011], using the freely available software STAN [Stan Development Team, 2014b], interfaced via the R package rstan [Stan Development Team, 2014a]. MCMC is an efficient computational technique to simulate random draws from an arbitrary (possibly unnormalized) probability distribution, such as our posterior distribution given by Eq. (9). A MCMC program can thus be regarded as a random number generator that samples from the posterior distribution. Using an appropriate MCMC sampler, we can approximate posterior distributions by smoothed histograms, and posterior expectations by empirical averages of samples drawn from the posterior distribution. The STAN software provides a scripting language to translate a user-specified generative model for the data (such as our signal-plus-noise model Eq. 1) into a MCMC sampler. The STAN model code for our analyses is provided in appendix A. The code shows that the derivation of the likelihood function Eq. (8) is not really required to implement the MCMC sampler in STAN – specification of the generative model Eq. (1) is enough. We have used the No-U-Turn sampler of STAN with all its

default settings. All our posterior distributions are based on 10^5 Monte-Carlo samples. These were generated by simulating 8 parallel Markov chains, each for 10^6 iterations, after discarding a warm-up period of 10^4 iterations for initialization of the algorithm. The 8 chains were thinned by retaining only every 80th sample to eliminate autocorrelation. Our procedure for generating the posterior samples takes about 20 minutes on a desktop computer with 8 CPUs. Reasonable results can, however, be obtained without thinning of the Markov chain, which reduces the time to generate 10^5 samples to a few seconds. Potential scale reduction factors close to one [Gelman and Rubin, 1992] and visual inspection of trace plots were taken as evidence for successful convergence and proper mixing of the Markov Chains.

2.4 Relation between ensemble mean and observations

The signal-plus-noise model can be used to learn about the relationship between the observations and the means of the ensemble forecasts. It follows from standard Normal theory [e.g., Mardia et al., 1979, sec. 3.2] that if the model parameters θ are known, the conditional distribution of the observation y_t given the ensemble mean $\bar{x}_t$ is

$$(y_t|\bar{x}_t, \theta) \sim \mathcal{N}\left(\mu_y + \frac{\beta \sigma_s^2}{\beta^2 \sigma_s^2 + \sigma_\eta^2/R}(\bar{x}_t - \mu_x), \sigma_\epsilon^2 + \sigma_s^2\left(\frac{\sigma_\eta^2}{R\beta^2\sigma_s^2 + \sigma_\eta^2}\right)\right). \quad (10)$$

In other words, the relationship between the observations y_t and ensemble means $\bar{x}_t$ is described by a simple linear regression model whose intercept, slope and residual variance parameters are functions of the known parameters of the signal-plus-noise model. So if there was no uncertainty in the signal-plus-noise parameters, one could use Eq. (10) as a basis for post-processing the ensemble means to predict the observations. Correcting dynamical forecasts by linear regression, also known as model output statistics [MOS, Glahn and Lowry, 1972], forms the basis for commonly used post-processing techniques in seasonal forecasting [e.g., Feddersen et al., 1999]. In sec. 3.6, we will compare the simple linear regression approach with a fully Bayesian posterior predictive approach which accounts for parameter uncertainty.

Eade et al. [2014] use the relation between signal-plus-noise interpretation and linear regression in their post-processing technique for the ensemble mean, and then adjust the distribution of the ensemble members around the post-processed ensemble mean to have the signal-to-noise ratio implied from the correlation, while retaining year-to-year variability in the ensemble spread. That is, while Eq. (10) assumes a constant variance, the method of Eade et al. [2014] allows for time varying ensemble variance. But Tippett et al. [2007] have shown for seasonal precipitation forecasts, that retaining the year-to-year variability of the ensemble variance does not improve the forecasts. The question of whether the ensemble spread should influence the width of the forecast distribution in seasonal NAO forecasting is not addressed further in this paper.
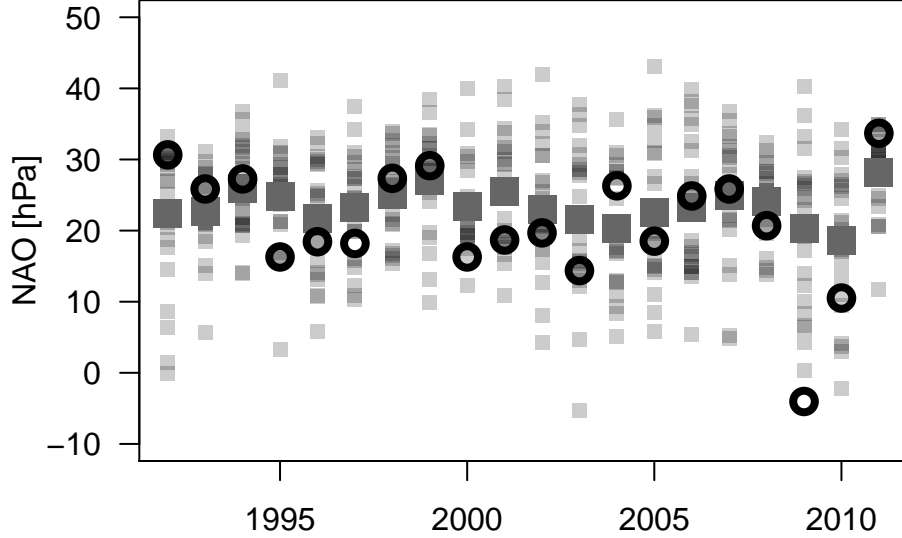
Figure 1: Raw winter NAO ensemble data generated by GloSea5 (semi-transparent squares), ensemble mean forecasts (large gray squares), and verifying NAO observations (circles).

3 Application to seasonal NAO hindcasts

3.1 The data

The signal-plus-noise model is demonstrated here by application to seasonal forecasts of the winter (Dec-Feb mean) North-Atlantic oscillation (NAO), discussed in Scaife et al. [2014]. Seasonal NAO predictability has further been studied by Doblas-Reyes et al. [2003], Eade et al. [2014], and Smith et al. [2014]. NAO is defined here as the difference in sea-level pressure between the Azores and Iceland (or nearest model grid points to these two locations; cf. Scaife et al. [2014]). A 24-member ensemble hindcast was generated annually from 1992 to 2011 by the Met Office Global Seasonal forecast System 5 (GloSea5), using lagged initialization between 25 October and 9 November (details about GloSea5 can be found in MacLachlan et al. [2014]). Raw forecast and observation data are shown in Fig. 1. In Table 1 we show a number of summary statistics of the hindcast data.

3.2 Prior specification

We use the following independent prior distribution functions for the model parameters: $\mu_x, \mu_y \sim \mathcal{N}(0, 30^2)$, $\sigma_s^2 \sim \mathcal{G}^{-1}(2, 25)$, $\sigma_\epsilon^2, \sigma_\eta^2 \sim \mathcal{G}^{-1}(3, 100)$, and $\beta \sim \mathcal{N}(1, 0.7^2)$, where $\mathcal{G}^{-1}(a, b)$ denotes the Inverse-Gamma distribution with shape parameter a and scale parameter b. A random variable $X \sim \mathcal{G}^{-1}(a, b)$

Table 1: Summary statistics of ensemble means $\bar{x}_t$ and observations y_t, and their particular values for the NAO hindcast.

$$
\begin{array}{rcll}
m_x & = & N^{-1}\sum_{t=1}^{N}\bar{x}_t & \quad 23.42 \quad hPa \\[4pt]
m_y & = & N^{-1}\sum_{t=1}^{N}y_t & \quad 20.94 \quad hPa \\[4pt]
v_{\bar{x}} & = & N^{-1}\sum_{t=1}^{N}(\bar{x}_t - m_x)^2 & \quad 5.24 \quad (hPa)^2 \\[4pt]
v_y & = & N^{-1}\sum_{t=1}^{N}(y_t - m_y)^2 & \quad 67.12 \quad (hPa)^2 \\[4pt]
s_{\bar{x}y} & = & N^{-1}\sum_{t=1}^{N}(\bar{x}_t - m_x)(y_t - m_y) & \quad 11.55 \quad (hPa)^2
\end{array}
$$

has a density function proportional to $x^{-a-1}\exp(-b/x)$. The Inverse-Gamma distribution was chosen as a prior because it is a common choice for variance parameters which can simplify Bayesian calculations. The prior distributions on μ_x and μ_y are very wide and uninformative, and we found the inference to be insensitive to the choice of these prior distributions. We found that the inference is more sensitive to the choice of priors on β and the σ parameters. These priors were deliberately chosen to be rather narrow: It can be shown by simulation experiments that, under the prior distributions above, σ_s has prior mean $\approx 4hPa$ and prior standard deviation of $\approx 2hPa$, and σ_η and σ_ϵ both have prior mean of $\approx 6.5hPa$ and prior standard deviation of $\approx 2.5hPa$. The parameters of the prior distributions were chosen by trial and error to yield reasonable prior distributions on observable quantities. In particular, the prior distributions of the standard deviation of the ensemble members ($\sqrt{var(x_i)}$, cf. Eq. 4b), and of the observation ($\sqrt{var(y)}$, cf. Eq. 4a) both have prior mean of $\approx 8hPa$ and prior standard deviation of $\approx 3hPa$. The correlation coefficient of the 24-member ensemble mean and the observations (cf. Eq. 6) has prior mean of ≈ 0.4 and prior standard deviation of ≈ 0.3, which covers sample correlation coefficients observed in past studies of seasonal winter NAO predictability (see, e.g., Kang et al. [2014] and Shi et al. [2015] for collections of seasonal winter NAO correlations obtained by different models). Furthermore, the prior probability of the model having lower signal-to-noise ratio than the observation is ≈ 0.5. The prior distributions on the model parameters therefore provide reasonable prior specifications for the analyses of sec. 3.4 (correlation coefficients) and sec. 3.5 (signal-to-noise ratios). It is worthwhile to point out that the priors are for horizontal atmospheric pressure differences measured in hPa; if NAO were measured differently, the above prior distributions would have to be rescaled.

The prior distribution is a subjective choice in Bayesian analysis, and is therefore often subject to criticism and discussion. We thus want to describe in more detail how we have arrived at the above distributions, and why we found default "uninformative" distributions unsatisfactory. We had initially specified independent uniform prior distributions on the model parameters as follows: $\sigma_{s,\epsilon,\eta} \sim U(0,30)$, and $\beta \sim U(-1,2)$ to cover physically meaningful ranges of the

parameter values, but without favoring a priori one set of parameters over the other. We have sampled model parameters from these prior distributions, and substituted the samples into the analytic expressions of the correlation coefficient given by Eq. (6). A smoothed histogram of the thus transformed samples approximates the derived prior distribution of the correlation coefficient. We found that the derived correlation prior has multiple modes, two of which are close to $+1$ and -1. Since the prior distribution should encode a priori information (for example about NAO prediction skill), this distribution is clearly unjustified. This example shows how seemingly "objective" and "uninformative" uniform prior distributions for the model parameters can lead to very informative and physically unjustified prior distributions on meaningful observable quantities. The uniform priors further produced a prior probability of over 0.6 that the model has a lower signal-to-noise ratio (SNR) than the observation. Since the possible anomalous signal-to-noise ratio in NAO predictions is a questions which we wanted to address, we did not want to bias the result a priori into the direction of a low model SNR. The chosen prior distributions represent a compromise between subjective judgements and previously published results about NAO variability, signal-to-noise ratio, and correlation skill.

In appendix D, the sensitivity to varying prior specification is illustrated for the correlation analysis of sec. 3.4. The analysis shows that while different priors indeed lead to different posteriors, the updated posterior distributions are similar. For more detailed discussions about the role and specification of prior distributions the reader is referred to the standard texts on Bayesian statistics given in sec. 2.3, in particular Gelman et al. [2004]. We lastly note that if sufficient data is available, the influence of the prior disappears and the Bayesian inference is dominated by the likelihood function [Gelman and Robert, 2013].

3.3 Bayesian updating

Having specified the prior distributions and the likelihood function, we now have all the ingredients to approximate the posterior distribution $p(\theta, s | x, y)$ by MCMC. Fig. 2 shows 200 MCMC samples, and the estimated posterior distributions of the parameters μ_x and μ_y. The posterior distributions of μ_x and μ_y were estimated from all 10^5 MCMC samples. The posterior distribution of μ_x is narrower than that of μ_y because the availability of 24 ensemble members allows for a more robust estimation of μ_x than μ_y, which is only based on one observational time series. Both posterior distributions, of μ_x and μ_y, have slightly heavier tails than the corresponding Normal distributions (not shown). The posterior means (standard deviations) are $23.4hPa$ ($0.56hPa$) for μ_x and $20.9hPa$ ($1.80hPa$) for μ_y. The model bias, defined by $\mu_x - \mu_y$ has posterior mean of $2.55hPa$ and posterior standard deviation of $1.64hPa$, resulting in a posterior probability of a positive bias $Pr(\mu_x > \mu_y) = 0.94$, and a posterior probability of 0.83 that the bias exceeds $1hPa$.

Fig. 3 shows that MCMC approximation allows for estimation of the latent variable s_t, of which 100 samples from the Markov Chain are shown (shifted upward by μ_y). The estimated time series of s_t are used in sec. 3.4 where we
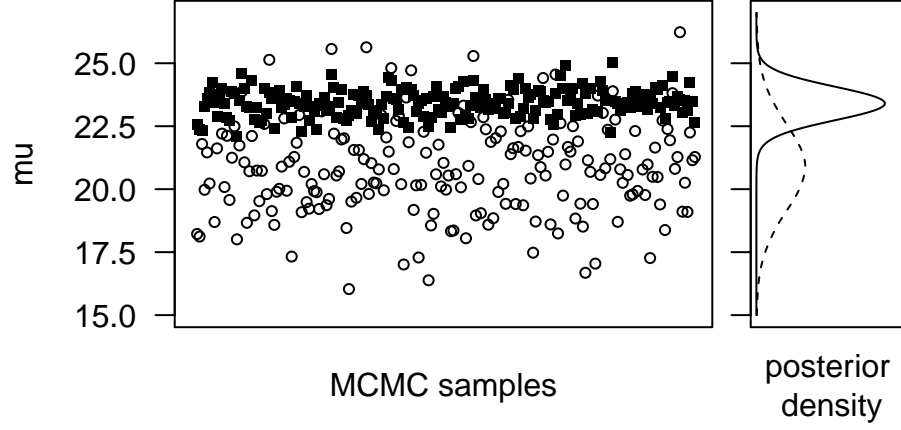
Figure 2: Illustration of MCMC approximation of posterior distributions. Left panel: Trace plot of 200 joint samples from the Markov Chain of μ_x (squares) and μ_y (circles). Right panel: Posterior distributions of μ_x (full line) and μ_y (dashed line), reconstructed from all 10^5 samples.
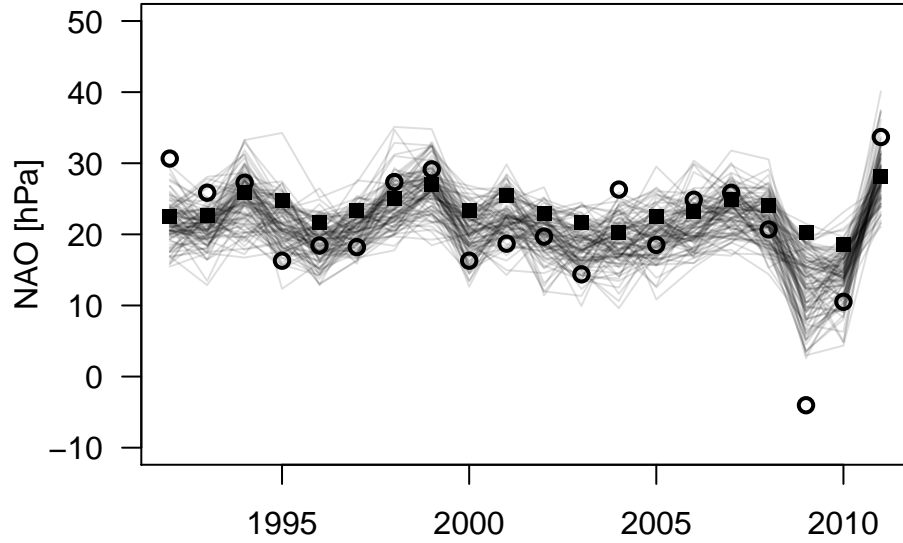

Figure 3: NAO observations (circles), GloSea5 ensemble means (squares) and 100 time-series of the variable $\mu_y + s_t$, drawn randomly from the Markov-Chain simulation (semi-transparent lines).
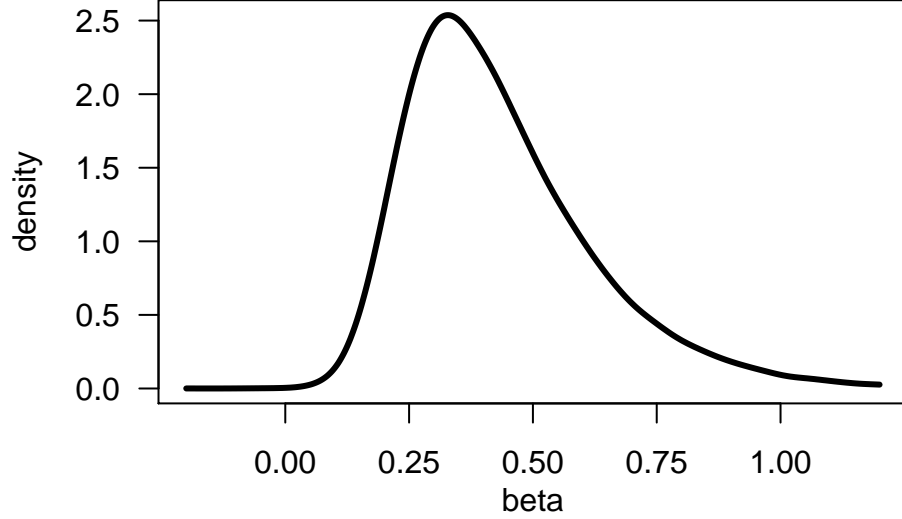
13

Figure 4: Posterior distribution of β.

generate new artificial ensemble forecasts for the 1992–2011 NAO observations to quantify uncertainty in correlation coefficients.

The posterior distribution for β is shown in Fig. 4. The parameter β quantifies how sensitive the forecasts are to the predictable signal relative to the sensitivity of the observations. When $\beta \neq 0$ there is dependency between forecast and observations - the forecasting system has skill. From the posterior distribution, the probability $Pr(\beta > 0) = 0.99$ and $Pr(\beta > 0.2) = 0.95$, and so we are confident that the forecasting system has skill for predicting the NAO. But are the forecasts reliable, i.e., are the raw ensemble members exchangeable with the observations? A necessary condition for reliability of the raw forecasts is that $\beta = 1$, which appears highly unlikely from our posterior distribution, which gives $Pr(\beta < 1) = 0.99$ and $Pr(\beta < 0.8) = 0.95$. This means that individual raw forecasts should not be taken on face value as possible realizations of the observations, which is in agreement with the conclusions of Eade et al. [2014], and highlights that statistical recalibration of the raw forecasts is necessary. Note that $\beta < 1$ implies that the model only contains a damped version of the predictable signal s_t. The posterior distribution of β thus indicates an anomalously low signal-to-noise ratio of the ensemble, which we analyze in more depth in sec. 3.5.

Fig. 5 shows the posterior distributions of the parameters σ_s, σ_ϵ, and σ_η. The posterior means (standard deviations) are $4.66 hPa$ $(1.53 hPa)$ for σ_s, $6.26 hPa$ $(1.22 hPa)$ for σ_ϵ, and $8.03 hPa$ $(0.26 hPa)$ for σ_η. It should be noted that σ_s and σ_ϵ are highly dependent: According to Eq. (4a), the sum of their squares is constrained by the variance of the observations; the total variance of the
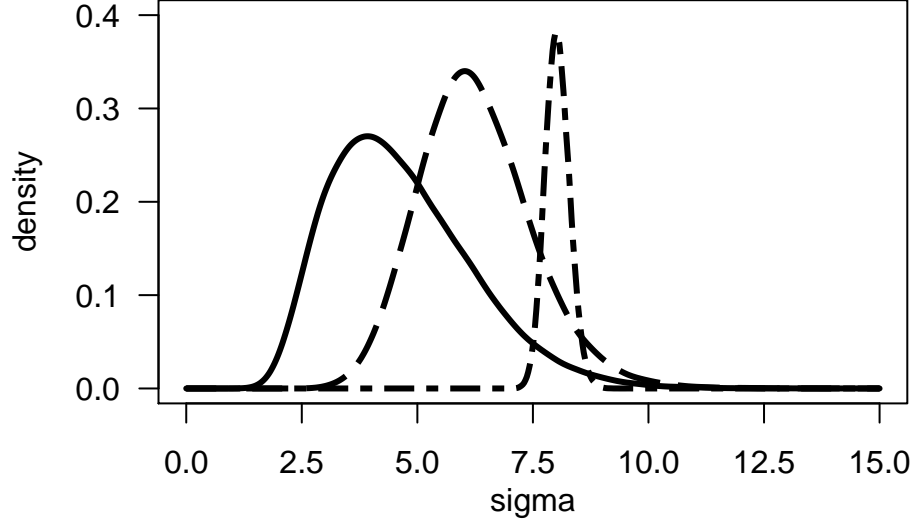
14

Figure 5: Marginal posterior distributions of σ_s (full line), σ_ϵ (dashed line), and σ_η (dot-dashed line; scaled by 1/4).

observations can be explained either by lots of signal and little noise, or little signal and lots of noise. If only the observations were available, σ_s and σ_ϵ would be unidentifiable. Only by basing the inference on the forecast system can σ_s (and therefore σ_ϵ) be constrained, however considerable uncertainty remains. In contrast to σ_ϵ, the parameter σ_η is better constrained by the data, because the individual ensemble members allow for estimation of the residual variance around the ensemble mean. A posterior comparison between the noise amplitudes yields $P(\sigma_\eta > \sigma_\epsilon) = 0.92$, i.e., there appears to be more unpredictable noise in the forecasting system than in the observations. At the same time, there is good agreement between the total standard deviations of the observations and the individual ensemble members, as defined in Eq. (4): The posterior mean (standard deviation) is $7.97 hPa$ $(1.09 hPa)$ for $\sqrt{var(y)}$, and $8.25 hPa$ $(0.28 hPa)$ for $\sqrt{var(x_i)}$.

Note that the model parameters are not invariant under linear transformations of either the observations or forecasts. However, since the NAO is often defined in different ways (e.g., by the leading sea level pressure empirical orthogonal function, or station pressure difference, or area averaged pressure difference, and possibly transformed to a normalized climate index), it is desirable that forecast performance should be based on quantities that are invariant to choice of linear scale. We will therefore now focus on scale-invariant functions of the parameters, namely the correlation coefficient in sec. 3.4, and signal-to-noise ratios in sec. 3.5.
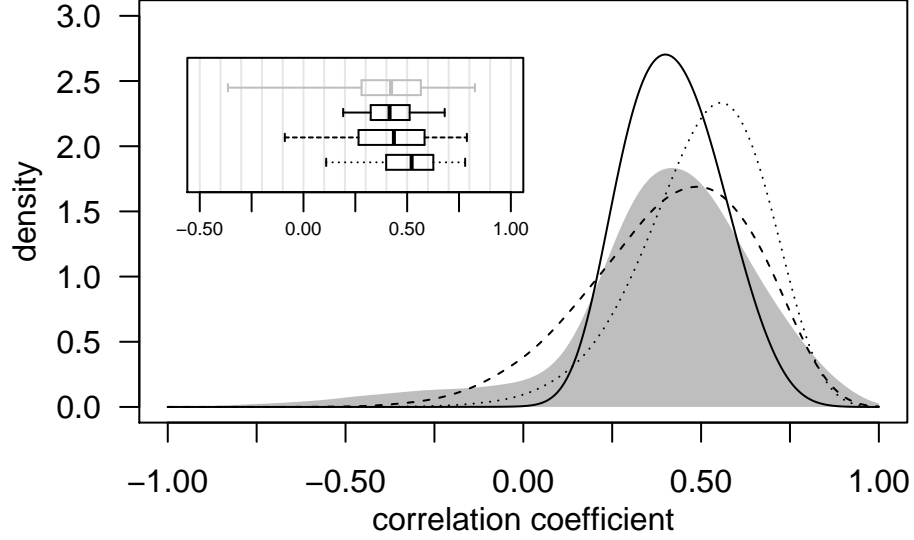
Figure 6: Uncertainty in the correlation coefficient. Prior of the population correlation ρ(gray area), its posterior distribution (solid line), the posterior predictive distribution of the sample correlation over arbitrary 20-year periods (dashed line), and the posterior predictive distribution of the sample correlation with observations fixed at the actual 1992–2011 NAO values (dotted line). The Box-and-Whiskers plots in the inset indicate the 2.5, 25, 50, 75, and 97.5 percentiles of the distributions.

3.4 Uncertainty in the correlation coefficient

A widely used evaluation criterion for ensemble mean forecasts is the Pearson correlation coefficient between the ensemble forecasts and observations given by

$$r_{\bar{x}y} = \frac{s_{\bar{x}y}}{\sqrt{v_{\bar{x}}v_y}}. \tag{11}$$

For the hindcast data presented in sec. 3.1, the sample correlation is $r_{\bar{x}y} = 0.62$. Uncertainty in correlation coefficients is usually quantified by confidence intervals and p-values [Von Storch and Zwiers, 2001, sec. 8.2.3]. This section presents a posterior analysis of uncertainty in the correlation coefficient of NAO hindcasts of sec. 3.1. We address three precise questions. It should be noted that the approach outlined below is applicable to other performance measures, such as the mean squared error (MSE), the continuous ranked probability score (CRPS), or the Ignorance score.

What is the uncertainty in the population correlation coefficient ρ, given the hindcast data? In other words: What are possible values of the

16

correlation coefficient taken over infinitely many 24-member ensembles and corresponding NAO observations, from which the given hindcast data is only a random sample of size $N = 20$? To answer this question, we consider the population correlation coefficient ρ of the 24-member ensemble mean, expressed as a function of the model parameters, as given by Eq. (6). We calculate ρ for each MCMC sample of the model parameters, and thereby approximate the posterior distribution of the population correlation coefficient. Our prior and updated posterior distribution of ρ are indicated by the gray area and the solid line in Fig. 6, respectively. The posterior distribution of ρ quantifies our uncertainty about the correlation coefficient due to uncertainty in the parameters of the statistical model, and due to the fact that the ensemble mean can only be estimated imperfectly by 24 ensemble members (thus the term σ_η^2/R in the denominator of Eq. 6). Due to the mode of the prior distribution of 0.4, the posterior mean of ρ of 0.42 is smaller than the actual sample correlation of 0.62. 20 samples of hindcast data are not sufficient to override the prior too much, and the result is therefore biased towards our prior judgments about NAO skill. It might be argued that this result is unduly influenced by the prior distribution on the correlation. In appendix D, we illustrate the sensitivity of the posterior distribution of the correlation for different prior distributions. The sensitivity analysis shows that even for a very optimistic prior distribution, with prior mode at a correlation of 0.7, the posterior mode of the correlation is shrunk down to about 0.5. The central 95% credible interval derived from the posterior distribution (depicted in the inset in Fig. 6 is equal to $[0.19, 0.68]$, which does not overlap zero, but which also has the sample correlation value of 0.62 in its upper tail. In conclusion, the sample correlation coefficient of 0.62 might be an overestimation of the true correlation skill of GloSea5, but we can say with high certainty that the system does have positive correlation skill.

What is the uncertainty in the sample correlation coefficient $r_{\bar{x}y}$ for different non-overlapping 20 year forecast periods? To answer this question, we calculate a large collection of sample correlation coefficients $r_{\bar{x}y}$ as follows: We draw a set of parameters $\{\mu_x, \mu_y, \beta, \sigma_s, \sigma_\epsilon, \sigma_\eta\}$ from the MCMC output. We use σ_s to sample a random signal time series $s_1, \cdots, s_{20}$, and then use the other parameters to generate a random hindcast data set with $R = 24$ and $N = 20$ according to Eq. (1). We then calculate the sample correlation $r_{\bar{x}y}$ of the ensemble mean in this artificial data set, and repeat this process for all 10^5 MCMC samples. The resulting distribution is the posterior predictive distribution of $r_{\bar{x}y}$ ("predictive" because it is a distribution over observables rather than parameters [Gelman et al., 2004]). The posterior predictive distribution is indicated by the dashed line in Fig. 6. This distribution accounts for parameter uncertainty (because we sample parameters from the posterior), and also for finite-sample uncertainty (because we draw a random hindcast data set of finite length N). The posterior predictive distribution therefore quantifies our uncertainty about the sample correlation calculated over an arbitrary 20-year period. The posterior mean and median of this predictive distribution are very

close to that of the posterior distribution of ρ. But the predictive distribution is wider than the posterior distribution. The 95% credible interval derived from this distribution is $[-0.09, 0.79]$. Taking into account finite sample uncertainty in addition to parameter uncertainty increases the overall uncertainty.

What is the uncertainty in the sample correlation coefficient $r_{\bar{x}y}$ for the same 1992–2011 NAO observations, but for a new realization of the ensemble forecast? To answer this question, we calculate the posterior predictive distribution of $r_{\bar{x}y}$, where the observations are fixed at their values shown in Fig. 1. That is, we generate replicated ensemble forecasts for these particular observations. To do this, we sample a signal time series $s_1, \cdots, s_{20}$ directly from the MCMC output (sketched in Fig. 3), instead of generating $s_1, \cdots, s_{20}$ randomly. We also draw the parameters β and σ_η from the same iteration of the Markov chain. We use these parameters to construct a new 24 member ensemble forecast using Eq. (1b), and then calculate the sample correlation with the original 1992-2011 NAO observations. Note that the sampled series of s_t is correlated with the original observations, and therefore the resampled ensemble members will be correlated with the original observations as well. The corresponding posterior predictive distribution of $r_{\bar{x}y}$ is indicated by the dotted line in Fig. 6. Treating the observations as fixed quantities and only the ensembles as random decreases the width of the distribution, the 95% credible interval is now $[0.11, 0.78]$. Furthermore, the predictive mean and mode of this distribution are about 0.5, i.e., slightly higher than the means and modes of the previous distributions. Our best explanation for this shift is that it is caused by the last 3 NAO observations which represent large excursions from the mean compared to the previous 17 observations, and thereby bias the correlation coefficient upward compared to randomly sampled observations from a Normal distribution. On the one hand this would imply that the Normal assumption is inadequate for the data. On the other hand, comparison of the two predictive distributions of $r_{\bar{x}y}$ (for fixed and arbitrary observational periods) suggests that in the future, when NAO might exhibit more normal behavior, the sample correlation using the same model will probably become smaller than 0.62.

3.5   Signal-to-noise analysis

It has been noted by Scaife et al. [2014], Kumar et al. [2014] and Eade et al. [2014] that the signal-to-noise ratios in seasonal climate predictions can be too low, which leads to the counter-intuitive effect that the ensemble forecasting system is less skillful at predicting members drawn from itself than at predicting the observation. This is problematic because the skill of the ensemble at predicting itself is often assumed to be an upper bound of predictability of the real world. Previous studies have provided only point estimates of signal-to-noise ratios and have not quantified how much uncertainty is in these quantities, which was criticized by Shi et al. [2015]. A Bayesian framework allows us to calculate posterior probabilities for hypotheses related to signal-to-noise ratios.
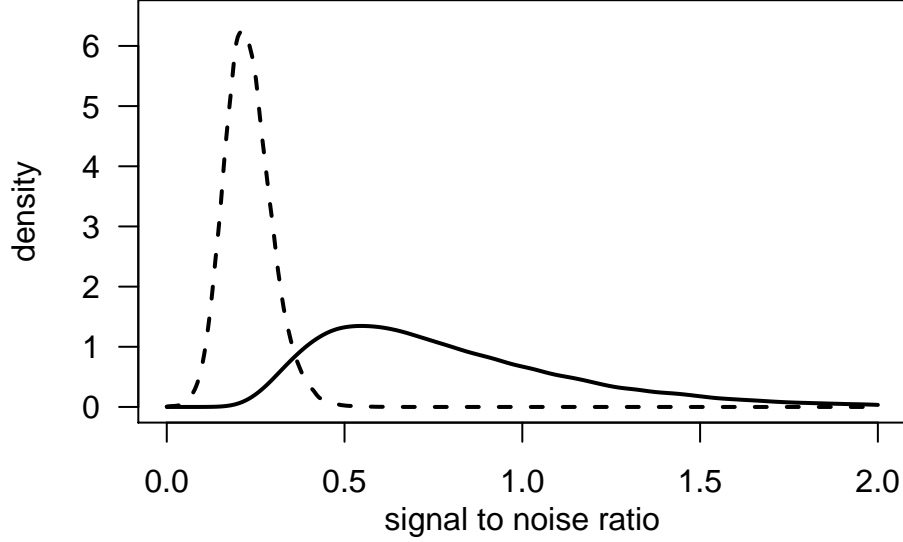
18

Figure 7: Posterior distributions of the signal-to-noise ratio of the observations (full line) and of the model (dashed line).

In Eade et al. [2014], the ratio of predictable components (RPC) was proposed as a measure to compare levels of predictability in the forecasting system and in the real world. The predictable component of the real world PC_{obs} was defined as the correlation between the ensemble mean and the observations, and the predictable component of the model PC_{mod} was defined as the ratio of the standard deviation of the ensemble mean and the mean standard deviation of the ensemble members. RPC equals the ratio PC_{obs}/PC_{mod}, and was found by Eade et al. [2014] to be about 2 for the NAO hindcast.

PC_{obs}, PC_{mod}, and RPC, expressed in terms of the parameters of the signal-plus-noise model are given in appendix B. Eade et al. [2014] argue that for a forecasting system that "perfectly reflects the actual predictability", RPC should be equal to one. If we define a perfect forecasting system by full exchangeability of ensemble members and observations, i.e., $\mu_x = \mu_y$, $\sigma_\epsilon = \sigma_\eta$, and $\beta = 1$, and substituting these equalities into Eq. (17c), we find that the perfect value of RPC is

$$\mathrm{RPC}_{\mathrm{perf}} = \left(1 + \frac{1}{R\sigma_s^2/\sigma_\epsilon^2}\right)^{-1}. \tag{12}$$

It can be noted from this that $\mathrm{RPC}_{\mathrm{perf}} \neq 1$ even for a fully exchangeable system. To obtain $\mathrm{RPC}_{\mathrm{perf}} = 1$, one also has to have either an infinitely large ensemble, i.e., $R \to \infty$, or no unpredictable noise in the system, i.e., $\sigma_\eta = \sigma_\epsilon = 0$. When both R and σ_ϵ are finite, RPC is smaller than one.

RPC is a rather complicated function of the $(\beta, \sigma_s, \sigma_\epsilon, \sigma_\eta)$ parameters (cf. Eq. 17c), and RPC = 1 corresponds to an imperfect forecasting system. There-

fore, we shall consider instead the signal-to-noise ratios (SNR) of the forecast system and of the observations. The SNR's are simply the ratio of the standard deviation of the predictable component (signal), and the unpredictable component (noise) of observations and of individual ensemble members, i.e.,

$$\text{SNR}_{\text{obs}} = \frac{\sigma_s}{\sigma_\epsilon}, \text{ and} \tag{13a}$$

$$\text{SNR}_{\text{mod}} = \frac{|\beta|\sigma_s}{\sigma_\eta}. \tag{13b}$$

Note that SNR_{obs} and SNR_{mod} are invariant under a shift or rescaling of the forecasts or the observations. Substituting the moment estimators from appendix C into Eq. (13), we obtain $\text{SNR}_{\text{obs}} = 1.73$ and $\text{SNR}_{\text{mod}} = 0.21$, i.e., the observations appear to be more predictable than the model. But the model parameters are very uncertain. Therefore we should also expect SNR's to be very uncertain.

Fig. 7 shows posterior distributions of SNR_{obs} and SNR_{mod} derived from the MCMC simulation. The posterior distribution of SNR_{mod} is sharper than that of SNR_{obs} because 24 ensemble members allow for more robust estimation than a single observation time series. We confirm with very high posterior probability the previous result of Scaife et al. [2014] that, for the GloSea5 winter NAO forecast, the SNR of the model is lower than the SNR of the observations. In particular, we have a posterior probability $Pr(\text{SNR}_{\text{obs}} > \text{SNR}_{\text{mod}}) = 0.99$ (updated from prior probability of ≈ 0.5). The sensitivity of this conclusion to the choice of the prior is briefly discussed in appendix D.

Our posterior analysis assigns very high probability to the hypothesis that the predictable signal component in the model is weaker than in the real world. The analysis of Shi et al. [2015], which is based on a set of winter NAO hindcasts produced by different models, concludes that such an underconfident ensemble "merely suggest an inadequately small sample size". Contrary to that, the analysis based on our 20-year data set (and our statistical assumptions) confirms the finding of Eade et al. [2014] with very high confidence, despite the small sample size: The raw GloSea5 ensemble underestimates the predictability of the real world, and statistical post-processing of the raw ensemble is necessary to generate reliable forecasts.

3.6 Calibration and prediction

Bayesian inference using the signal-plus-noise model provides a natural framework for recalibrating forecasts to produce reliable probability distributions of future observations. The predictive distribution function for the unknown observation y_t is the conditional distribution of y_t, given the known quantities $\{x, y\}_{-t}$, i.e., the hindcast data set not including the time instance t, as well as x_t, i.e., the ensemble forecast for y_t. The predictive distribution can be calculated by integrating over the posterior distribution of the model parameters:

$$p(y_t|\{x, y\}_{-t}, x_t) = \int d\theta \; p(y_t|x_t, \theta)p(\theta|\{x, y\}_{-t}, x_t). \tag{14}$$

20

Note that according to Eq. (10), the conditional distribution $p(y_t|x_t, \theta)$ is a Normal distribution whose parameters depend on the signal-plus-noise model parameters. The predictive distribution Eq. (14) can thus be interpreted as a weighted mixture of Normal distributions, where the weight is given by the posterior density of the model parameters. A mixture of Normal distributions is itself not in general a Normal distribution. We should thus not expect the predictive distributions to be Normal, even though our statistical model is based on the assumption of Normality of the data. The resulting predictive distributions include a suitable predictive variance that takes into account parameter uncertainties and forecast uncertainty. We generate $N = 20$ predictive distributions in leave-one-out mode, that is for each $t = 1, \cdots, N$, the predictive distribution for y_t is calculated under the assumption that y_t is unknown (see Friedman et al. [2009] for further details on cross validation). The STAN code has to be adjusted slightly to simulate these out-of-sample predictive distributions (see appendix A).

We compare the posterior predictive distribution functions to a simple benchmark given by ordinary linear regression. Recall that we have argued in sec. 2.4 that if the model parameters were known, linear regression would be the optimal post-processing method for signal-plus-noise models. We regress the observations y_t on the ensemble means $\bar{x}_t$, and predict a Gaussian forecast distribution with the residual variance of the regression, i.e.,

$$(y_t|\bar{x}_t) \sim \mathcal{N}\left(m_y + \frac{s_{\bar{x}y}}{v_{\bar{x}}}(\bar{x}_t - m_x), v_y\left(1 - r_{\bar{x}y}^2\right)\right) \qquad (15)$$

The benchmark predictions were generated in leave-one-out mode as well.

The posterior predictive distributions and the benchmark predictions are shown in Fig. 8. In general the posterior predictive distributions are wider than the benchmark predictions; their average standard deviations are $7.5hPa$ and $6.6hPa$, respectively. The posterior predictive means are less variable than the benchmark means; their standard deviations are $3.0hPa$ and $5.1hPa$, respectively.

The larger dispersion of the posterior predictive distributions leads to the effect that in the majority of cases (when the NAO is close to its climatological mean) the benchmark predictions assign a higher predictive density to the observation than the posterior predictive distributions. On the other hand, if the observation is far away from the climatological mean, or far away from the forecast mean, the posterior predictive distributions assign more density to the observations. We address the question of which collection of forecasts are "better" on average by calculating the average Ignorance score [Roulston and Smith, 2002]. Given a forecast density $p(z)$ and a verifying observation y, the Ignorance score is defined by

$$\mathcal{I}(p; y) = -\log_2 p(y). \qquad (16)$$

The Ignorance is a proper scoring rule for probabilistic forecasts of continuous quantities; its average can be taken as a summary of forecast performance, indicating better forecasts by lower values.
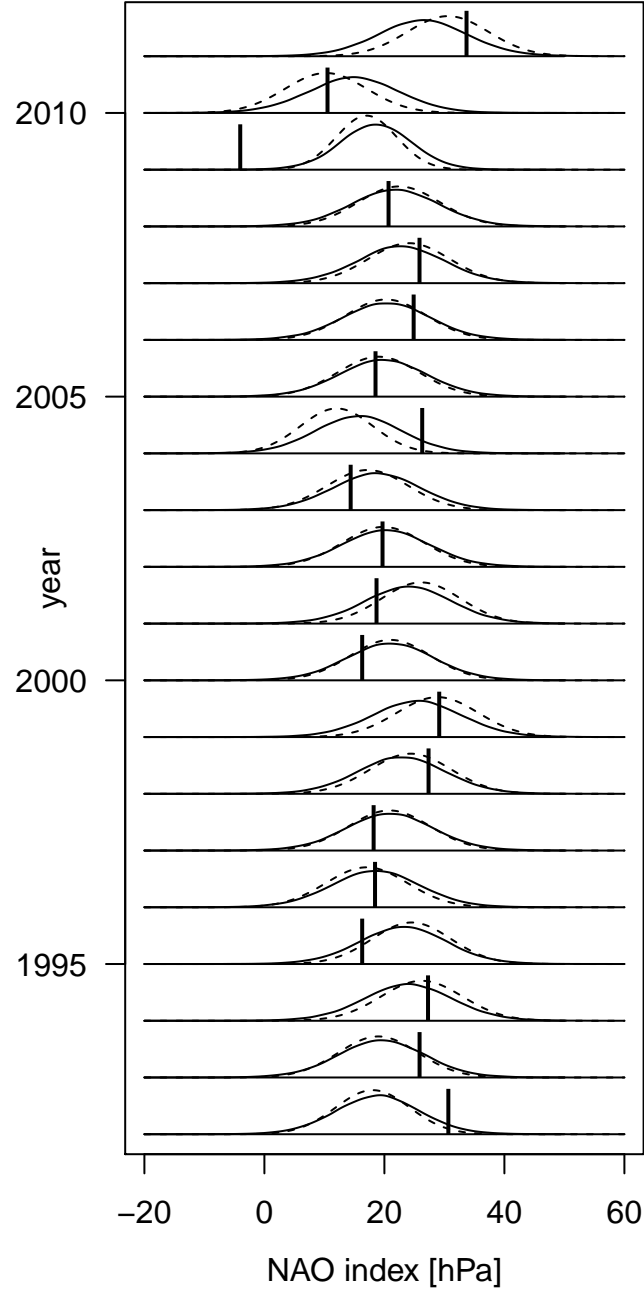
Figure 8: Predictive distributions functions based on simple linear regression (dashed line) and fully Bayesian, using the signal-plus-noise model (full lines). The vertical lines indicate the observations.

Table 2: Average ignorance scores and standard errors for different forecast methods

Method	mean Ign.	standard error
climatology	5.46	0.62
regression benchmark	5.24	0.62
posterior predictive	5.02	0.41

In Table 2 we compare the average Ignorance scores of three different forecasts: The leave-one-out climatological forecast which is simply a Normal distribution with the climatological mean and variance, the linear regression benchmark, and the posterior predictive distributions. It is reassuring that the posterior predictive distributions assign a higher average density to the observation than both, the climatology and the regression benchmark. The additional skill is due to the wider predictive distributions and the less variable predictive mean. These two features are a consequence of accounting for parameter uncertainty by integrating over their posterior distribution. In conclusion, the Bayesian analysis using a signal-plus-noise model not only provides useful evaluation diagnostics, but also provides a natural way of generating skillful and well-calibrated probability forecasts.

4 Discussion

4.1 Model criticism

We have used a simplified statistical model to make inference about an actual forecasting system, so it is important to be aware of the limitations of the statistical model. It is important not to confuse limitations of our statistical model with deficiencies of the real forecasting system.

There are a number of features of observed climate indices and their ensemble forecasts that our simplified model cannot account for. These include: Autocorrelation in the ensemble forecasting system and the observations; a spread-skill relation, that is, a systematic relationship between the ensemble spread and the distance between the ensemble mean and the verifying observation; trend in the observations and drifts in the model output; skewness, bimodality, or heavy-tailedness of the distribution of the predictand. More work is necessary to develop statistical frameworks for ensemble forecasts that take some or all of these effects into account without becoming overly complex. On the other hand, by leaving out all these details, our model retains a high level of interpretability. Before making the model more complex we also have to ask ourselves, how much information can we justifiably hope to infer from 20 years' worth of annual hindcast data?

4.2 Model checking

We have tested the validity of our exchangeability assumptions in sec. 2.1 by replacing the observation by one of the ensemble members. Since we judged the ensemble members to be exchangeable, replacing the observation by an ensemble member should produce a perfect model scenario, where the observation and ensemble members are statistically indistinguishable from each other, i.e., we should have $\mu_x = \mu_y$, $\beta = 1$, $\sigma_\epsilon = \sigma_\eta$. After rerunning the posterior analysis under this perfect model scenario, we found that the posterior distributions of μ_x and μ_y, and of σ_ϵ and σ_η overlap each other and provide no indication for non-exchangeability. Furthermore, the posterior distribution of β does not rule out the value $\beta = 1$ as strongly as the posterior distribution shown in Fig. 4. However, we still found the bulk of the posterior distribution of β to be concentrated between 0 and 1, resulting in a rather high posterior probability of $Pr(\beta < 1) \approx 0.95$. Furthermore, we found a posterior probability for an anomalous signal-to-noise ratio of $Pr(\text{SNR}_{\text{perf. obs}} > \text{SNR}_{\text{mod}}) \approx 0.85$ in this perfect-model scenario. These posterior probabilities provide evidence that our statistical model might not be flexible enough to accurately model the data. A possible explanation for the observed behavior is that the ensemble members are, in fact, not exchangeable with each other, which could be the result of the lagged initialization of GloSea5 on 3 different dates. Without further analyses we are unable to model such an ensemble with non-exchangeable members and we leave this problem open for future studies.

We have also repeated our analyses with different NAO observations, taken directly from station data at Lisbon and Reykjavik [Hurrell, James and National Center for Atmospheric Research Staff (Eds), 2014], and from the leading empirical orthogonal function of sea level pressure anomalies over the Atlantic sector [National Center for Atmospheric Research Staff (Eds), 2014]. The posterior distributions of the μ's and σ's change slightly because the alternative observations have different scales. For the scale-invariant quantities analyzed in sec. 3.4 and 3.5, however, the posterior distributions are almost identical to the ones we obtained earlier. Our main conclusions are therefore insensitive to the choice of NAO observations.

4.3 Correlation uncertainty under different hindcast settings

Statistical inference using a signal-plus-noise model might be useful for design of future ensemble systems. Simulations from the model can be used to calculate predictive distributions of correlation coefficients for different ensemble sizes R and for different sample sizes N. In practice, the hindcast length N and the ensemble size R can usually not be chosen independently, but their choice is constrained by the available computational resources. Given that the computational expense of a planned hindcast experiment, defined by the product NR, is fixed, how should N and R be chosen? One possible criterion might be to consider the range of possible values of the correlation coefficient. Fig. 9 shows
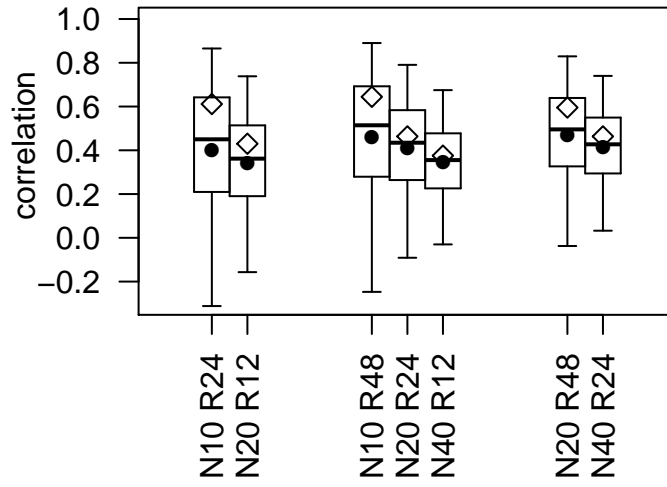
Figure 9: Posterior predictive distribution of correlation coefficients for different combinations of N and R. The box-and-whiskers indicate the 2.5, 25, 50, 75, and 97.5 percentiles of the predictive distributions, the diamonds indicate the mode and the circles the mean. The plots are grouped according to their computational expense $N \times R$.

that for a given computational expense (i.e., NR constant), there is a trade off between mean and spread of the distribution of possible correlation values. Higher expected correlation can be obtained by increasing the ensemble size R while decreasing the hindcast length N. At the same time, however, the risk of getting very low sample correlations (e.g., not significantly different from 0) increases if N is decreased. This is because the spread of possible correlation values becomes wider, but also because the larger N is, the smaller will be the correlation values that are deemed "significant" by statistical tests.

5   Conclusion

This study has shown how a statistical model can be used to diagnose and improve the skill and reliability of an ensemble forecasting system. The distributions-oriented approach [Murphy and Winkler, 1987] provides a complete summary of the forecasting system and observations using a signal-plus-noise model, whose parameters can be estimated by Bayesian inference. Posterior distributions of the parameters can be used to simulate properties of any desired performance measure and its uncertainty under hypothetical designs of the ensemble forecasting system. The framework provides a straightforward method for calculating calibrated probability forecasts for future observables for a given set of ensemble forecasts.

We conclude by revisiting the 5 questions specified in the introduction which guided the analysis of NAO hindcasts produced by the GloSea5 seasonal climate prediction system. Question 1: There is indeed much sampling uncertainty in the correlation between the ensemble mean and observations. But there is also strong evidence of actual positive correlation skill: the 95% credible interval of $[0.19, 0.68]$ does not overlap zero. Question 2: Our analysis suggests that very different correlation skill might be observed over different 20 year periods. In particular, the value of 0.62 is in the upper tail of the correlation distribution, suggesting a high chance of a decrease in correlation skill if GloSea5 were evaluated over different periods. Question 3: The skill uncertainty over the same 1992–2010 evaluation period is smaller than over arbitrary 20 year evaluation periods. Our results suggest that the 20 year period is unusual and produces higher than normal correlation skill. The reasons for this are not entirely clear, but might be related to large deviations of NAO in the years 2008–2010. Question 4: Forecasts are certainly not exchangeable with the observations and can therefore benefit from recalibration. A particular feature of non-exchangeability is the anomalous signal-to-noise ratio (SNR). We show with over 99% posterior probability that the SNR is smaller in the model than in the observations, i.e. the predictable signal in the model is too weak. Question 5: The probabilistic framework used in this study allows us to derive a recalibrated predictive distribution, i.e. the conditional distribution of the observation, given the ensemble forecast. We found that the Bayesian method of integrating over the parameter uncertainty distribution improves the forecast skill compared to a simpler recalibration method.

It is worthwhile to highlight a few important advantages of a Bayesian framework over more traditional approaches. Firstly, the proposed statistical model is based on explicit assumptions, which creates transparency in how we interpret the observed data, and about how we think forecasts are related to the real world. Transparency is the basis for critically discussing assumptions, and revising these assumptions if necessary. Secondly, all the analyses to answer our research questions are coherently based on the exact same assumptions about the data. There are established methods to address each of our research questions in isolation, for example a t-test for the correlation coefficient [Von Storch and Zwiers, 2001], analysis of ratio of predictable components [RPC, Eade et al., 2014] to address signal-to-noise ratio, and non-homogeneous Gaussian regression [NGR, Gneiting et al., 2005] for forecast recalibration. But these methods are not explicitly based on the same statistical assumptions. An explicit statistical model allows us to address different questions in a coherent way, without changing our assumptions about the data. Lastly, uncertainty quantification is a crucial aspect of analyzing small climate hindcast data sets. In Bayesian analyses, probability is the primitive quantity, and uncertainty quantification is therefore built into the analysis by default. All questions can be addressed by posterior probability distributions, which not only communicate our best guesses, but also our degree of uncertainty. On the other hand, computational methods for Bayesian analyses can be expensive, the specification of suitable prior distributions is problematic, and all conclusions are conditional on the parametric model assumptions being correct.

In future studies, it will be of interest to relax model assumptions (e.g., to include serial dependence in the signal time series), and to extend the model to allow for possible sources of non-stationarity (e.g., climate change trends), as well as spread-skill relationships. A more disciplined way of specifying the prior distribution over model parameters is needed. It will also be of interest to develop computationally efficient methods for modeling spatial ensemble hindcasts and observations available at many grid point locations.

Acknowledgments

Appendix A: STAN model code

For the diagnostic analysis, where all N observations and ensemble forecasts are known, the following STAN code was used to approximate the posterior

distribution:

```
data {
  int<lower=1> N;
  int<lower=1> R;
  matrix[N,R] x;
  vector[N] y;
}
parameters {
  real mu_x;
  real mu_y;
  real<lower=0> sigma2_eps;
  real<lower=0> sigma2_eta;
  real<lower=0> sigma2_s;
  real beta;
  vector[N] s;
}
model {
  mu_x ~ normal(0, 30);
  mu_y ~ normal(0, 30);
  beta ~ normal(1, 0.7);
  sigma2_s ~ inv_gamma(2, 25);
  sigma2_eps ~ inv_gamma(3, 100);
  sigma2_eta ~ inv_gamma(3, 100);
  s ~ normal(0, sqrt(sigma2_s));

  y ~ normal(mu_y + s, sqrt(sigma2_eps));
  for (n in 1:N)
    for (r in 1:R)
      x[n,r] ~ normal(mu_x + beta * s[n], sqrt(sigma2_eta));
}
```

In order to generate the predictive distributions for sec. 3.6, where the N-th observation is assumed to be unknown, the following STAN code was used:

```
data {
  int<lower=1> N;
  int<lower=1> R;
  matrix[N,R] x;
  vector[N-1] y;
}
parameters {
  real mu_x;
  real mu_y;
  real<lower=0> sigma2_eps;
  real<lower=0> sigma2_eta;
  real<lower=0> sigma2_s;
```

```
    real beta;
    vector[N-1] s;
    real s_new;
}
model {
  mu_x ~ normal(0, 30);
  mu_y ~ normal(0, 30);
  beta ~ normal(1, 0.7);
  sigma2_s ~ inv_gamma(2, 25);
  sigma2_eps ~ inv_gamma(3, 100);
  sigma2_eta ~ inv_gamma(3, 100);
  s ~ normal(0, sqrt(sigma2_s));

  y ~ normal(mu_y + s, sqrt(sigma2_eps));
  for (n in 1:(N-1))
    for (r in 1:R)
      x[n,r] ~ normal(mu_x + beta * s[n], sqrt(sigma2_eta));

  s_new ~ normal(0, sigma_s);
  for (r in 1:R)
      x[N,r] ~ normal(mu_x + beta * s_new, sqrt(sigma2_eta));
}
generated quantities {
  real y_new;
  y_new <- normal_rng(mu_y + s_new, sqrt(sigma2_eps));
}
```

Appendix B: Ratio of predictable components as functions of model parameters

This appendix complements sec. 3.5. PC_{obs}, PC_{mod}, and RPC, expressed in terms of the parameters of the signal-plus-noise model are given by

$$PC_{obs} = \frac{\beta\sigma_s^2}{\sqrt{(\sigma_s^2 + \sigma_\epsilon^2)(\beta^2\sigma_s^2 + \sigma_\eta^2/R)}}, \tag{17a}$$

$$PC_{mod} = \sqrt{\frac{\beta^2\sigma_s^2 + \sigma_\eta^2/R}{\beta^2\sigma_s^2 + \sigma_\eta^2}}, \text{ and} \tag{17b}$$

$$RPC = \sqrt{\frac{1 + \sigma_\eta^2/(\beta^2\sigma_s^2)}{\left(1 + \sigma_\epsilon^2/\sigma_s^2\right)\left(1 + \sigma_\eta^2/(R\beta^2\sigma_s^2)\right)}}. \tag{17c}$$

29

Appendix C: Method of moment estimators for the signal-plus-noise model

To calculate moment estimators for the parameters of the signal-plus-noise model Eq. (1), we use the summary measures given in Table 1, and additionally the average ensemble variance

$$v_x = (NR)^{-1} \sum_{t=1}^{N} \sum_{r=1}^{R} (x_{t,r} - \bar{x}_t)^2. \tag{18}$$

Equating the analytical first and second moments (cf. Eq. 4) of the signal-plus-noise model with sample moments, and solving for the model parameters, we obtain estimating equations for the model parameters. The equations and corresponding values for the NAO data of sec. 3.1 are summarised in Table 3.

Table 3: Estimating equations for parameters in the signal-plus-noise model derived by method of moments, and estimated values for the GloSea5 NAO hindcast data.

Estimating equations			Values	
$\hat{\mu}_x$	$=$	m_x	23.42	hPa
$\hat{\mu}_y$	$=$	m_y	20.94	hPa
$\hat{\sigma}_\eta^2$	$=$	v_x	62.17	$(hPa)^2$
$\hat{\beta}$	$=$	$s_{\bar{x}y}^{-1}(v_{\bar{x}} - R^{-1}v_x)$	0.23	
$\hat{\sigma}_s^2$	$=$	$\hat{\beta}^{-1} s_{\bar{x}y}$	50.35	$(hPa)^2$
$\hat{\sigma}_\epsilon^2$	$=$	$v_y - \hat{\sigma}_s^2$	16.77	$(hPa)^2$

Appendix D: Sensitivity to choice of priors

Bayesian analyses are sensitive to the choice of prior distributions. This is desired for the present study; we want our prior judgments about diagnostic quantities to have an impact on our conclusions, especially because the sample size is small. In order to illustrate the sensitivity to the choice of prior distributions, we consider the variability of the posterior distribution of the correlation coefficient ρ (cf. sec. 3.4 and Eq. 6) when the prior parameters are varied. We found the shape of the prior distribution on ρ to be sensitive to the shape and scale parameters of the Inverse-Gamma prior distribution on σ_s^2. We have varied these parameters within values that produce believable prior distributions on ρ. We have then calculated new posterior distributions of ρ using the alternative prior specifications. The varying prior distributions and updated posterior distributions of ρ are shown in Fig. 10. As expected, due to the small sample size the posterior distributions vary considerably due to the variability of the prior.
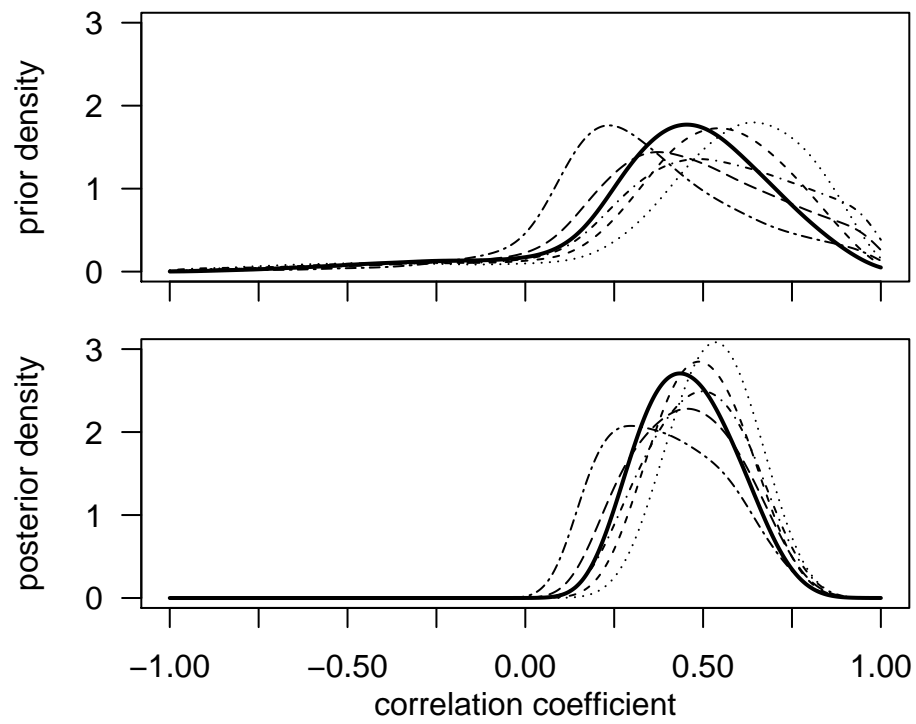
Figure 10: Different prior distributions on the correlation coefficient (top panel) yield different posterior distributions (lower panel, with corresponding line types). The bold solid lines corresponds to the specifications of sec. 3.3
.

But the differences between the different prior distributions is greater than the differences between their updated posterior distributions. Bayesian updating leads to a consensus between differing prior judgments. Note further that the optimistic prior distributions with prior mode at ≈ 0.7 is shrunk towards a mode at ≈ 0.5, which is smaller than the sample correlation of 0.62 for the data.

In sec. 3.5 we have shown that there is a high posterior probability of an anomalously low signal-to-noise ratio of the model: $Pr(\text{SNR}_{\text{obs}} > \text{SNR}_{\text{mod}}) > 0.99$. This probability is sensitive to the choice of the prior parameters. Note that by changing only the prior distribution of σ_s^2, the prior distribution of the correlation changes, but the prior probability $Pr(\text{SNR}_{\text{obs}} > \text{SNR}_{\text{mod}}) \approx 0.5$ does not change. We found that by changing the prior of σ_s^2 in such a way that the correlation prior becomes more pessimistic, the posterior probability for $(\text{SNR}_{\text{obs}} > \text{SNR}_{\text{mod}})$ decreases. For the prior specifications that yield the most pessimistic prior expectation of the correlation of ≈ 0.3 in Fig. 10, the posterior probability of an anomalous SNR reduces to ≈ 0.85.

References

J. Annan and J. Hargreaves. Reliability of the CMIP3 ensemble. *Geophysical Research Letters*, 37(2), 2010.

D. J. Bartholomew, M. Knott, and I. Moustaki. *Latent Variable Models and Factor Analysis: A Unified Approach*, volume 899. John Wiley & Sons, 2011.

A. A. Bradley, S. S. Schwartz, and T. Hashino. Distributions-oriented verification of ensemble streamflow predictions. *Journal of Hydrometeorology*, 5(3): 532–545, 2004.

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.

J. P. Buonaccorsi. *Measurement Error: Models, Methods, and Applications*. CRC Press, 2010.

R. E. Chandler. Exploiting strength, discounting weakness: Combining information from multiple climate simulators. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371 (1991), 2013.

F. Doblas-Reyes, V. Pavan, and D. Stephenson. The skill of multi-model seasonal forecasts of the wintertime North Atlantic Oscillation. *Climate dynamics*, 21(5-6):501–514, 2003.

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

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. CRC press, 1994.

B. S. Everitt. *An Introduction to Latent Variable Models*. Springer, 1984.

H. Feddersen, A. Navarra, and M. N. Ward. Reduction of model systematic error by statistical correction for dynamical seasonal predictions. *J. Climate*, 12(7):1974–1989, Jul 1999.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, volume 2. Springer, Berlin, 2009. URL `http://statweb.stanford.edu/~tibs/ElemStatLearn/`.

W. A. Fuller. *Measurement Error Models*. John Wiley & Sons, 1987.

A. Gelman and C. P. Robert. "Not Only Defended But Also Applied": The Perceived Absurdity of Bayesian Inference. *The American Statistician*, 67(1): 1–5, Feb 2013.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman Hall/CRC, 2004.

H. R. Glahn and D. A. Lowry. The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology*, 11(8): 1203–1211, Dec 1972.

T. Gneiting, A. E. Raftery, A. H. Westveld III, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.

L. Goddard, A. Kumar, A. Solomon, D. Smith, G. Boer, P. Gonzalez, V. Kharin, W. Merryfield, C. Deser, S. J. Mason, et al. A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics*, 40(1-2): 245–272, 2013.

Hurrell, James and National Center for Atmospheric Research Staff (Eds). The Climate Data Guide: Hurrell North Atlantic Oscillation (NAO) Index (station-based). Retrieved online from https://climatedataguide.ucar.edu/climate-data/hurrell-north-atlantic-oscillation-nao-index-station-based [last modified 05 Sep 2014], 2014.

E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge university press, 2003.

I. T. Jolliffe and D. B. Stephenson. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons, 2012.

D. Kang, M.-I. Lee, J. Im, D. Kim, H.-M. Kim, H.-S. Kang, S. D. Schubert, A. Arribas, and C. MacLachlan. Prediction of the arctic oscillation in boreal winter by dynamical seasonal forecasting systems. *Geophys. Res. Lett.*, 41 (10):3577–3585, May 2014.

V. V. Kharin and F. W. Zwiers. Improved seasonal probability forecasts. *Journal of Climate*, 16(11):1684–1701, 2003.

A. Kumar, P. Peng, and M. Chen. Is there a relationship between potential and actual skill? *Monthly Weather Review*, 142(6):2220–2227, 2014.

D. V. Lindley. *Understanding uncertainty*. John Wiley & Sons, 2006.

C. MacLachlan, A. Arribas, K. A. Peterson, A. Maidens, D. Fereday, A. A. Scaife, M. Gordon, M. Vellinga, A. Williams, R. E. Comer, and et al. Global Seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, Jun 2014.

R. A. Madden. Estimates of the natural variability of time-averaged sea-level pressure. *Monthly Weather Review*, 104(7):942–952, 1976.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic press, 1979.

P. Moran. Estimating structural and functional relationships. *Journal of Multivariate Analysis*, 1(2):232–255, 1971.

A. H. Murphy and D. S. Wilks. A case study of the use of statistical models in forecast verification: Precipitation probability forecasts. *Weather and Forecasting*, 13(3):795–810, 1998.

A. H. Murphy and R. L. Winkler. A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338, 1987.

J. M. Murphy. Assessment of the practical utility of extended range ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 116(491): 89–125, Jan 1990. ISSN 1477-870X. doi: 10.1002/qj.49711649105. URL `http://dx.doi.org/10.1002/qj.49711649105`.

National Center for Atmospheric Research Staff (Eds). The Climate Data Guide: Hurrell North Atlantic Oscillation (NAO) Index (PC-based). Retrieved online from https://climatedataguide.ucar.edu/climate-data/hurrell-north-atlantic-oscillation-nao-index-pc-based [last modified 05 Sep 2014], 2014.

F. E. Otto, C. Ferro, T. Fricker, and E. Suckling. On judging the credibility of climate projections. *Climatic Change*, 2012.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge Univ Press, 2000.

E. E. Riddle, A. H. Butler, J. C. Furtado, J. L. Cohen, and A. Kumar. Cfsv2 ensemble prediction of the wintertime arctic oscillation. *Climate Dynamics*, 41(3-4):1099–1116, Jul 2013.

C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation (Springer Texts in Statistics)*. Springer New York, 2007.

J. Rougier, M. Goldstein, and L. House. Second-Order Exchangeability Analysis for Multimodel Ensembles. *Journal of the American Statistical Association*, 108(503):852–863, 2013.

M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6), 2002.

P. G. Sansom, D. B. Stephenson, C. A. Ferro, G. Zappa, and L. Shaffrey. Simple uncertainty frameworks for selecting weighting schemes and interpreting multimodel ensemble climate change experiments. *Journal of Climate*, 26(12): 4017–4037, 2013.

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

W. Shi, N. Schaller, D. MacLeod, T. Palmer, and A. Weisheimer. Impact of hindcast length on estimates of seasonal climate predictability. *Geophys. Res. Lett.*, Feb 2015.

D. M. Smith, A. A. Scaife, R. Eade, and J. R. Knight. Seasonal to decadal prediction of the winter North Atlantic Oscillation: Emerging capability and future prospects. *Quarterly Journal of the Royal Meteorological Society*, 2014.

Stan Development Team. RStan: the R interface to Stan, Version 2.5.0, 2014a. URL `http://mc-stan.org/rstan.html`.

Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual, Version 2.5.0*, 2014b. URL `http://mc-stan.org/`.

D. B. Stephenson, M. Collins, J. C. Rougier, and R. E. Chandler. Statistical problems in the probabilistic prediction of climate change. *Environmetrics*, 23(5):364–372, 2012.

C. Tebaldi, R. L. Smith, D. Nychka, and L. O. Mearns. Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *Journal of Climate*, 18(10):1524–1540, 2005.

M. K. Tippett, A. G. Barnston, and A. W. Robertson. Estimation of seasonal precipitation tercile-based categorical probabilities from ensembles. *J. Climate*, 20(10):2210–2228, May 2007.

H. Von Storch and F. W. Zwiers. *Statistical analysis in climate research.* Cambridge University Press, 2001.

X. Wang and C. H. Bishop. Improvement of ensemble reliability with a new dressing kernel. *Quarterly Journal of the Royal Meteorological Society*, 131 (607):965–986, 2005.

A. P. Weigel, M. A. Liniger, and C. Appenzeller. Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? *Monthly Weather Review*, 137(4):1460–1479, 2009.