A Bayesian Framework for Verification and Recalibration of Ensemble Forecasts: How Uncertain is NAO Predictability?

Stefan Siegert, David B. Stephenson, and Philip G. Sansom

University of Exeter, Exeter, United Kingdom

Adam A. Scaife, Rosie Eade, and Alberto Arribas

Met Office Hadley Centre, Exeter, United Kingdom

(Manuscript received 11 March 2015, in final form 30 July 2015)

ABSTRACT

Predictability estimates of ensemble prediction systems are uncertain because of 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 Global Seasonal Forecast System, version 5 (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 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 the following:

1) Could the observed skill be due to chance sampling: that is, natural variability in the observed events and ensemble of forecasts?
2) How might the skill vary for a different nonoverlapping 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; that is, do the individual model forecasts have similar properties to the observations?
5) How can nonexchangeable ensemble forecasts be used to create a reliable probability forecast of future observations?

Corresponding author address: Stefan Siegert, University of Exeter, Laver Building, North Park Road, Exeter EX4 4QE, United Kingdom.
E-mail: s.siegert@exeter.ac.uk

DOI: 10.1175/JCLI-D-15-0196.1

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 nonhomogeneous 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 modeling 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 datasets 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 dataset of past forecasts and observations and are therefore uncertain. We therefore advocate a framework that 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 dataset.

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, and 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). In section 2, we discuss these in more detail, 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 section 3, the proposed statistical framework is used to analyze recent North Atlantic Oscillation (NAO) hindcasts made with the Met Office Global Seasonal Forecast System, version 5 (GloSea5; MacLachlan et al. 2014; Scaife et al. 2014). We demonstrate how the framework allows us to coherently address questions 1–5 above: that is, to analyze uncertainty in correlation skill, assess the exchangeability of forecasts and observations, and transform raw ensemble forecasts into recalibrated predictive distribution functions.

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 because of model errors, internal variability, measurement error, etc. (the noise). Although the same or similar statistical models have been used in previous studies (summarized in section 2b), we will provide a detailed discussion of the underlying statistical assumptions and their implications.

a. The signal-plus-noise model

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

$$y_t = \mu_y + s_t + \varepsilon_t \quad \text{and} \tag{1a}$$

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

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 ($\beta \neq 1$ and $\sigma_\varepsilon \neq \sigma_\eta$).

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:

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

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

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

The $(R + 1) \times (R + 1)$ dimension covariance matrix $\boldsymbol{\Sigma}$ has the following entries:

$$\mathrm{var}(y) = \sigma_s^2 + \sigma_\varepsilon^2, \tag{4a}$$

$$\mathrm{var}(x_i) = \beta^2\sigma_s^2 + \sigma_\eta^2, \tag{4b}$$

$$\mathrm{cov}(x_i, x_j) = \beta^2\sigma_s^2 (i \neq j), \quad \text{and} \tag{4c}$$

$$\mathrm{cov}(x_i, y) = \beta\sigma_s^2, \tag{4d}$$

for all $i, j = 1, \ldots, R$. Therefore, the model Eq. (1) can be considered as a simplified parameterization 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 $\boldsymbol{\Sigma}$ 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

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

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

$$\rho = \frac{\mathrm{cov}(\overline{x}, y)}{[\mathrm{var}(\overline{x})\mathrm{var}(y)]^{1/2}} = \frac{\beta\sigma_s^2}{[(\beta^2\sigma_s^2 + \sigma_\eta^2/R)(\sigma_s^2 + \sigma_\varepsilon^2)]^{1/2}}. \tag{6}$$

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 $\sigma_\varepsilon = \sigma_\eta$. 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 $\sigma_\varepsilon = \sigma_\eta = 0$ 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 $\beta^2\sigma_s^2 + \sigma_\eta^2 = \sigma_s^2 + \sigma_\varepsilon^2$.

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, section 17.2.2; 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.

b. *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 section 3e, 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 model to infer future climate from a collection of exchangeable climate model runs. Chandler (2013) provides a statistical framework for multimodel ensembles, where runs from one climate model are nonexchangeable with runs from different climate models and nonexchangeable 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 model with $\beta = 1$, $\mu_x = \mu_y$, and $\sigma_\varepsilon = \sigma_\eta$.

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:

$$y_t = s_t + \varepsilon_t \quad \text{and} \tag{7a}$$

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

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

c. *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 complications can arise because of negative variance estimates that 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: that is, 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 (e.g., the posterior mean or mode) and uncertainty intervals (e.g., the 95% parameter values with highest posterior density). We denote $\theta = \{\mu_x, \mu_y, \beta, \sigma_s, \sigma_\varepsilon, \sigma_\eta\}$, the collection of unknown parameters of the signal-plus-noise model; $s = \{s_1, \ldots, s_N\}$, the unknown values of the latent signal variable; and $\{x, y\} = \{x_{t,1}, \ldots, x_{t,R}, y_t\}_{t=1}^N$, the collection of known forecasts and observations from a hindcast experiment. The desired posterior distribution for Bayesian estimation is thus $p(\theta, s \mid x, y)$. Derivation of the posterior distribution requires the specification of a prior probability distribution $\pi(\theta, s)$ over the unknown quantities, which factors into $\pi(\theta)\prod_{t=1}^N \pi(s_t \mid \sigma_s)$ 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 section 3b). 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 $\ell(x, y \mid \theta, s)$, can be calculated from Eq. (1) using the distribution law of the normal distribution, and the independence assumption:

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

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

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

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 Stan software (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). An 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 an 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 eight parallel Markov chains, each for 10^6 iterations, after discarding a warm-up period of 10^4 iterations for initialization of the algorithm. The eight chains were thinned by retaining only every 80th sample to eliminate autocorrelation. Our procedure for generating the posterior samples takes about 20 min on a desktop computer with eight 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.

d. 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, their section 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

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

In other words, the relationship between the observations y_t and ensemble means $\bar{x}_t$ is described by a simple linear regression model for which the 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 postprocessing the ensemble means to predict the observations. Correcting dynamical forecasts by linear regression, also known as MOS (Glahn and Lowry 1972), forms the basis for commonly used postprocessing techniques in seasonal forecasting (e.g., Feddersen et al. 1999). In section 3f, we will compare the simple linear regression approach with a fully Bayesian posterior predictive approach that accounts for parameter uncertainty.

Eade et al. (2014) use the relation between signal-plus-noise interpretation and linear regression in their postprocessing technique for the ensemble mean and then adjust the distribution of the ensemble members around the postprocessed 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.

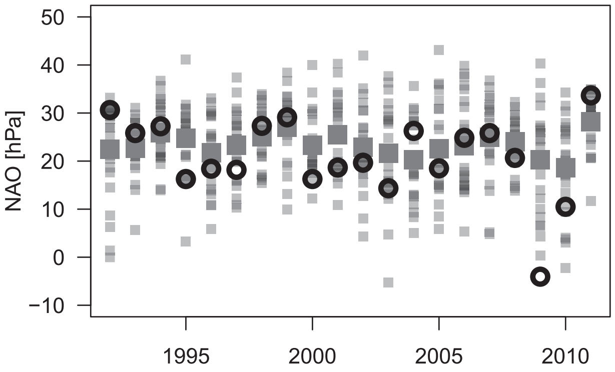

FIG. 1. Raw winter NAO ensemble data generated by GloSea5 (small gray-shaded squares), ensemble mean forecasts (large gray squares), and verifying NAO observations (circles).

3. Application to seasonal NAO hindcasts

a. The data

The signal-plus-noise model is demonstrated here by application to seasonal forecasts of the winter (December–February mean) 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 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. (Data used to generate figures, graphs, plots, and tables are freely available via contacting the lead author at s.siegert@exeter.ac.uk.)

b. Prior specification

We use the following independent prior distribution functions for the model parameters: μ_x, $\mu_y \sim \mathcal{N}(0, 30^2)$, $\sigma_s^2 \sim \mathcal{G}^{-1}(2, 25)$, σ_ε^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 $X \sim \mathcal{G}^{-1}(a, b)$ 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 that can simplify Bayesian calculations. The prior distributions on μ_x and μ_y are very wide and uninformative, and we found

TABLE 1. Summary statistics of ensemble means $\overline{x}_t$ and observations y_t and their particular values for the NAO hindcast.

Summary statistics	Values
$m_x = N^{-1}\sum_{t=1}^{N}\overline{x}_t$	23.42 hPa
$m_y = N^{-1}\sum_{t=1}^{N} y_t$	20.94 hPa
$v_{\overline{x}} = N^{-1}\sum_{t=1}^{N}(\overline{x}_t - m_x)^2$	5.24 hPa^2
$v_y = N^{-1}\sum_{t=1}^{N}(y_t - m_y)^2$	67.12 hPa^2
$s_{\overline{x}y} = N^{-1}\sum_{t=1}^{N}(\overline{x}_t - m_x)(y_t - m_y)$	11.55 hPa^2

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$4 hPA and prior standard deviation of approximately 2 hPa, and σ_η and σ_ε both have prior mean of approximately 6.5 hPa and prior standard deviation of approximately 2.5 hPa. 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 $\{[\mathrm{var}(x_i)]^{1/2}$, cf. Eq. (4b)$\}$ and of the observation $\{[\mathrm{var}(y)]^{1/2}$, cf. Eq. (4a)$\}$ both have prior mean of approximately 8 hPa and prior standard deviation of approximately 3 hPa. The correlation coefficient of the 24-member ensemble mean and the observations [cf. Eq. (6)] has prior mean of approximately 0.4 and prior standard deviation of approximately 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 approximately 0.5. The prior distributions on the model parameters therefore provide reasonable prior specifications for the analyses of section 3d (correlation coefficients) and section 3e (signal-to-noise ratios). It is worthwhile to point out that the priors are for horizontal atmospheric pressure differences measured in hectopascals; 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: $\sigma_{s,\varepsilon,\eta} \sim U(0, 30)$ 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 (e.g., 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 question 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 section 3d. 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 section 2c, in particular Gelman et al. (2004). We last 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).

c. Bayesian updating

Having specified the prior distributions and the likelihood function, we now have all the ingredients to approximate the posterior distribution $p(\theta, s \mid x, y)$ by MCMC. Figure 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

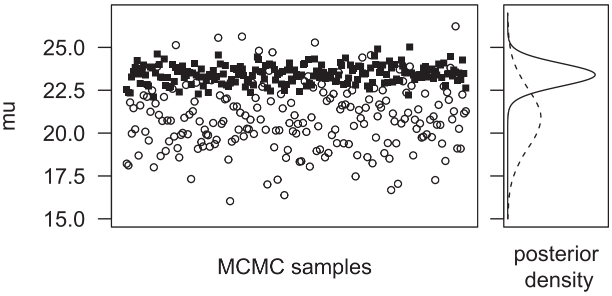

FIG. 2. Illustration of MCMC approximation of posterior distributions. (left) Trace plot of 200 joint samples from the Markov chain of μ_x (filled squares) and μ_y (circles). (right) Posterior distributions of μ_x (solid line) and μ_y (dashed line), reconstructed from all 10^5 samples.

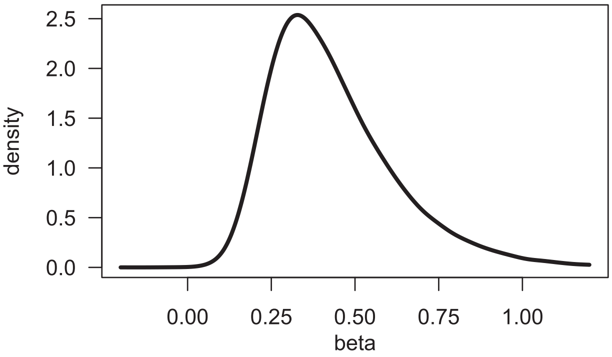

FIG. 4. Posterior distribution of β.

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 23.4 hPa (0.56 hPa) for μ_x and 20.9 hPa (1.80 hPa) for μ_y. The model bias, defined by $\mu_x - \mu_y$ has posterior mean of 2.55 hPa and posterior standard deviation of 1.64 hPa, 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 1 hPa.

Figure 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 section 3d, where we 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 $\Pr(\beta > 0.2) = 0.95$ so we are confident that the forecasting system has skill for predicting the NAO. But are the forecasts reliable; that is, 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 at 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 section 3e.

Figure 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

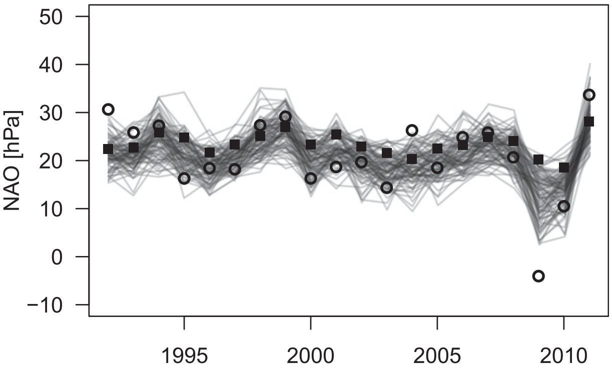

FIG. 3. NAO observations (circles), GloSea5 ensemble means (filled squares), and 100 time series of the variable $\mu_y + s_t$, drawn randomly from the Markov chain simulation (light gray lines).

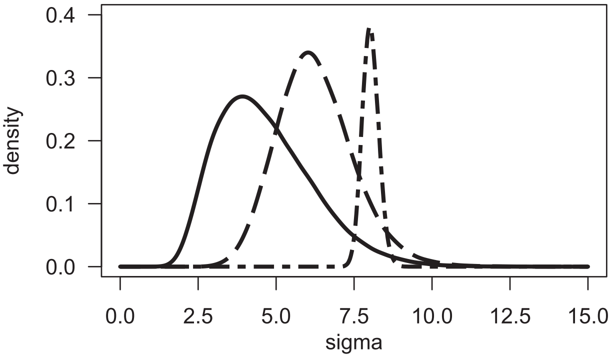

FIG. 5. Marginal posterior distributions of σ_s (solid line), σ_ε (dashed line), and σ_η (dotted–dashed line; scaled by 1/4).

(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 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 $\Pr(\sigma_\eta > \sigma_\varepsilon) = 0.92$; that is, 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 $[\mathrm{var}(y)]^{1/2}$, and 8.25 hPa (0.28 hPa) for $[\mathrm{var}(x_i)]^{1/2}$.

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, 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 section 3d and signal-to-noise ratios in section 3e.

d. 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

$$r_{\overline{x}y} = \frac{s_{\overline{x}y}}{(v_{\overline{x}}v_y)^{1/2}}. \tag{11}$$

For the hindcast data presented in section 3a, the sample correlation is $r_{\overline{x}y} = 0.62$. Uncertainty in correlation coefficients is usually quantified by confidence intervals and p values (Von Storch and Zwiers 2001, section 8.2.3). This section presents a posterior analysis of uncertainty in the correlation coefficient of NAO hindcasts of section 3a. 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.

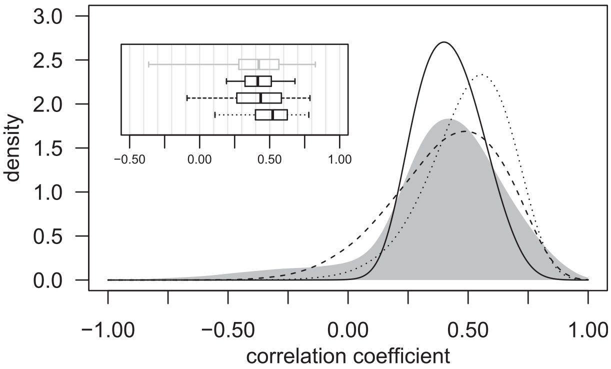

FIG. 6. Uncertainty in the correlation coefficient. Prior of the population correlation ρ (gray area), its posterior distribution (solid line), the posterior predictive distribution of the sample correlation over arbitrary 20-yr periods (dashed line), and the posterior predictive distribution of the sample correlation with observations fixed at the actual 1992–2011 NAO values (dotted line). The box-and-whisker plots in the inset indicate the 2.5, 25, 50, 75, and 97.5 percentiles of the distributions.

(i) What is the uncertainty in the population correlation coefficient ρ, given the hindcast data? In other words, what are possible values of the correlation coefficient taken over infinitely many 24-member ensembles and corresponding NAO observations, from which the given hindcast data are 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 distributions 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)]. As a result of 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. Since 20 samples of hindcast data are not sufficient to override the prior too much, the result is biased toward 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.

(ii) What is the uncertainty in the sample correlation coefficient $r_{\overline{x}y}$ for different nonoverlapping 20-yr forecast periods? To answer this question, we calculate a large collection of sample correlation coefficients $r_{\overline{x}y}$ as follows: We draw a set of parameters $\{\mu_x, \mu_y, \beta, \sigma_s, \sigma_\varepsilon, \sigma_\eta\}$ from the MCMC output. We use σ_s to sample a random signal time series $s_1, \ldots, s_{20}$ and then use the other parameters to generate a random hindcast dataset with $R = 24$ and $N = 20$ according to Eq. (1). We then calculate the sample correlation $r_{\overline{x}y}$ of the ensemble mean in this artificial dataset and repeat this process for all 10^5 MCMC samples. The resulting distribution is the posterior predictive distribution of $r_{\overline{x}y}$ [“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 dataset of finite length N). The posterior predictive distribution therefore quantifies our uncertainty about the sample correlation calculated over an arbitrary 20-yr 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.

(iii) What is the uncertainty in the sample correlation coefficient $r_{\overline{x}y}$ 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 $r_{\overline{x}y}$, 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 $s_1, \ldots, s_{20}$ directly from the MCMC output (sketched in Fig. 3), instead of generating $s_1, \ldots, s_{20}$ 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 $r_{\overline{x}y}$ 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 three 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 $r_{\overline{x}y}$ (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.

e. *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 counterintuitive 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.

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., $\mu_x = \mu_y$, $\sigma_\varepsilon = \sigma_\eta$, and $\beta = 1$) and substitute these equalities into Eq. (B1c), we find that the perfect value of RPC is

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

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., $\sigma_\eta = \sigma_\varepsilon = 0$). When both R and σ_ε are finite, $\mathrm{RPC}_{\mathrm{perf}}$ is smaller than one.

RPC is a rather complicated function of the parameters β, σ_s, σ_ε, and σ_η [cf. Eq. (B1c)], and RPC = 1 corresponds to an imperfect forecasting system. Therefore, we shall consider instead the SNRs of the forecast system and of the observations. The SNRs 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; that is,

$$\mathrm{SNR}_{\mathrm{obs}} = \frac{\sigma_s}{\sigma_\varepsilon} \quad \text{and} \qquad (13a)$$

$$\mathrm{SNR}_{\mathrm{mod}} = \frac{|\beta|\sigma_s}{\sigma_\eta}. \qquad (13b)$$

Note that $\mathrm{SNR}_{\mathrm{obs}}$ and $\mathrm{SNR}_{\mathrm{mod}}$ 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 $\mathrm{SNR}_{\mathrm{obs}} = 1.73$ and $\mathrm{SNR}_{\mathrm{mod}} = 0.21$ (i.e., the observations appear to be more predictable than the model). But the model parameters are very uncertain. Therefore, we should also expect SNRs to be very uncertain.

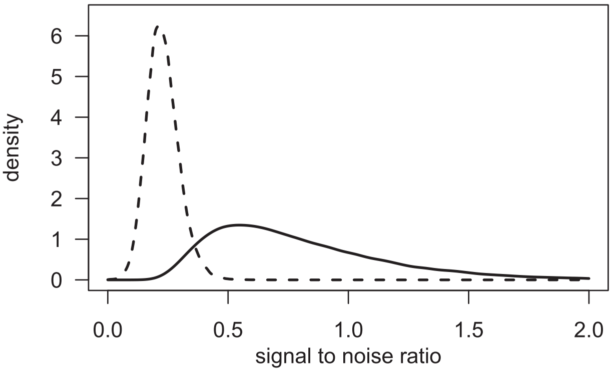

FIG. 7. Posterior distributions of the signal-to-noise ratio of the observations (solid line) and of the model (dashed line).

Figure 7 shows posterior distributions of $\mathrm{SNR}_{\mathrm{obs}}$ and $\mathrm{SNR}_{\mathrm{mod}}$ derived from the MCMC simulation. The posterior distribution of $\mathrm{SNR}_{\mathrm{mod}}$ is sharper than that of $\mathrm{SNR}_{\mathrm{obs}}$ 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 $\Pr(\mathrm{SNR}_{\mathrm{obs}} > \mathrm{SNR}_{\mathrm{mod}}) = 0.99$ (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-yr dataset (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 postprocessing of the raw ensemble is necessary to generate reliable forecasts.

f. 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 dataset 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:

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

Note that, according to Eq. (10), the conditional distribution $p(y_t \mid x_t, \theta)$ is a normal distribution, the parameters of which 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 $t = 1, \ldots, N$, the predictive distribution for y_t is calculated under the assumption that y_t is unknown [see Hastie 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 section 2d that, if the model parameters were known, linear regression would be the optimal postprocessing 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; that is,

$$(y_t \mid \bar{x}_t) \sim \mathcal{N}\left[m_y + \frac{s_{\bar{x}y}}{v_{\bar{x}}}(\bar{x}_t - m_x), v_y(1 - r^2_{\bar{x}y})\right]. \tag{15}$$

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

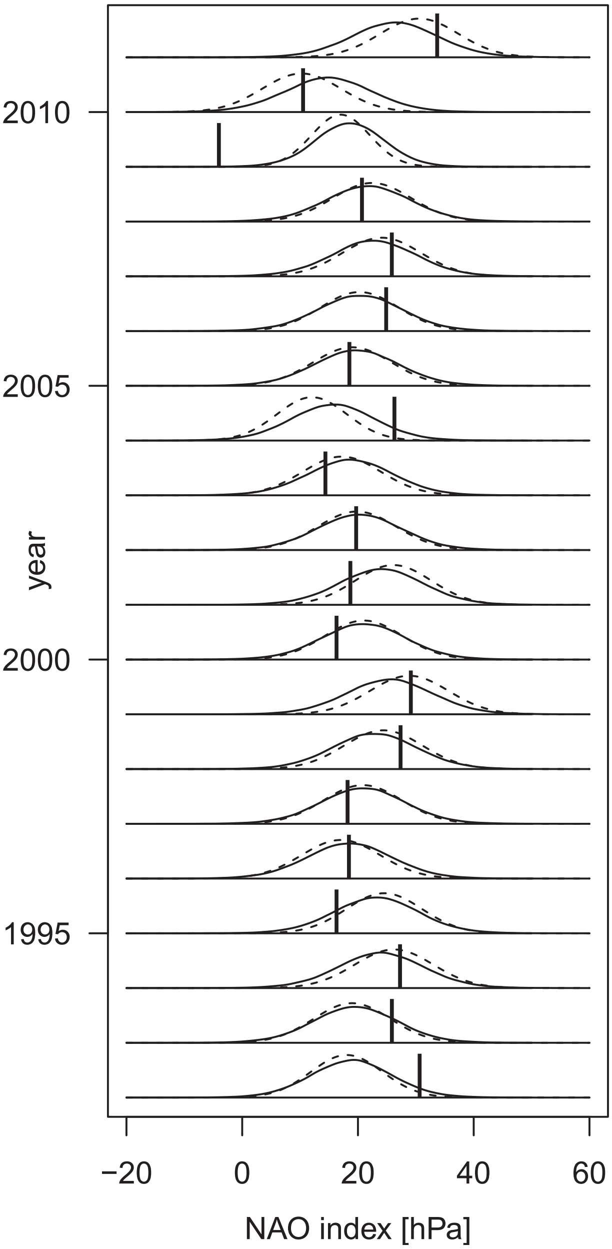

FIG. 8. Predictive distributions functions based on simple linear regression (dashed line) and fully Bayesian, using the signal-plus-noise model (solid lines). The vertical lines indicate the observations.

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 7.5 and 6.6 hPa, respectively. The posterior predictive means are less variable than the benchmark means; their standard deviations are 3.0 and 5.1 hPa, 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

TABLE 2. Average ignorance scores and standard errors for different forecast methods.

Method	Mean ignorance	Standard error
Climatology	5.46	0.62
Regression benchmark	5.24	0.62
Posterior predictive	5.02	0.41

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 is 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

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

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.

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

a. Model criticism

We have used a simplified statistical model to make inferences 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 for which our simplified model cannot account. These include the following: 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; and 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?

b. Model checking

We have tested the validity of our exchangeability assumptions in section 2a by replacing the observation with one of the ensemble members. Since we judged the ensemble members to be exchangeable, replacing the observation with 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 $\mu_x = \mu_y$, $\beta = 1$, and $\sigma_\varepsilon = \sigma_\eta$). 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 nonexchangeability. 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 $\Pr(\mathrm{SNR}_{\mathrm{perf.obs}} > \mathrm{SNR}_{\mathrm{mod}}) \approx 0.85$ 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 three different dates. Without further analyses, we are unable to model such an ensemble with nonexchangeable 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, Portugal, and Reykjavik, Iceland, (NCAR staff 2014a) and from the leading empirical orthogonal function of sea level pressure anomalies over the

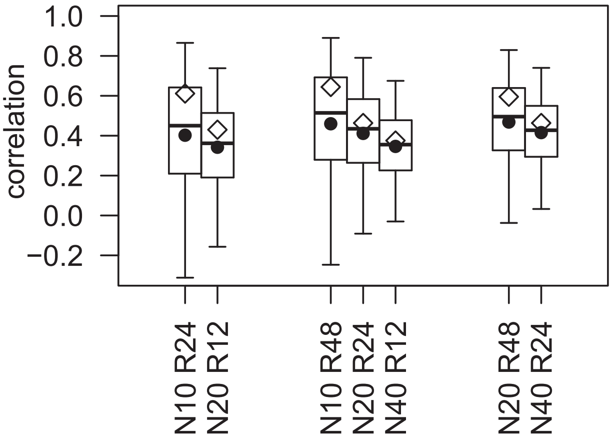

FIG. 9. Posterior predictive distribution of correlation coefficients for different combinations of N and R. The box-and-whisker plots indicate the 2.5, 25, 50, 75, and 97.5 percentiles of the predictive distributions; the diamonds indicate the mode and the dots indicate the mean. The plots are grouped according to their computational expense $N \times R$.

Atlantic sector (NCAR staff 2014b). The posterior distributions of μ and σ change slightly because the alternative observations have different scales. For the scale-invariant quantities analyzed in sections 3d and 3e, 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.

c. *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. Figure 9 shows 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. Conclusions

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, the parameters of which 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 five 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-yr 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-yr evaluation periods. Our results suggest that the 20-yr 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 2008–10. Question 4: Forecasts are certainly not exchangeable with the observations and can therefore benefit from recalibration. A particular feature of nonexchangeability 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: that is, 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. First, 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. Second, 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 Student's *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 nonhomogeneous 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. Last, uncertainty quantification is a crucial aspect of analyzing small climate hindcast datasets. 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 nonstationarity (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 gridpoint locations.

Acknowledgments. We wish to thank Theo Economou, Ben Youngman, and Danny Williamson for valuable discussions on Bayesian theory and MCMC. Further comments from Peter Watson and three anonymous reviewers greatly improved the paper. This work was supported by the European Union Programme FP7/2007-13 under Grant Agreement 3038378 (SPECS) and by the joint DECC/Defra Met Office Hadley Centre Climate Programme (GA1101).

APPENDIX A

Stan Model Code

For the diagnostic analysis, where all *N* observations and ensemble forecasts are known, the following algorithm in 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));
 }
```

To generate the predictive distributions for section 3f, where the *N*th observation is assumed to be unknown, the following algorithm in 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 section 3e. $\mathrm{PC}_{\mathrm{obs}}$, $\mathrm{PC}_{\mathrm{mod}}$, and RPC, expressed in terms of the parameters of the signal-plus-noise model are given by

$$\mathrm{PC}_{\mathrm{obs}} = \frac{\beta\sigma_s^2}{[(\sigma_s^2+\sigma_\varepsilon^2)(\beta^2\sigma_s^2+\sigma_\eta^2/R)]^{1/2}}, \tag{B1a}$$

$$\mathrm{PC}_{\mathrm{mod}} = \left(\frac{\beta^2\sigma_s^2+\sigma_\eta^2/R}{\beta^2\sigma_s^2+\sigma_\eta^2}\right)^{1/2}, \quad \text{and} \tag{B1b}$$

$$\mathrm{RPC} = \left\{\frac{1+\sigma_\eta^2/(\beta^2\sigma_s^2)}{(1+\sigma_\varepsilon^2/\sigma_s^2)[1+\sigma_\eta^2/(R\beta^2\sigma_s^2)]}\right\}^{1/2}. \tag{B1c}$$

TABLE C1. 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
$\sigma_\varepsilon^2 = v_y - \hat{\sigma}_s^2$	16.77 hPa^2

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:

$$v_x = (NR)^{-1}\sum_{t=1}^{N}\sum_{r=1}^{R}(x_{t,r}-\bar{x}_t)^2. \tag{C1}$$

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 section 3a are summarized in Table C1.

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. To illustrate the sensitivity to the choice of prior distributions, we consider the variability of the posterior distribution of the correlation coefficient ρ [cf. section 3d 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

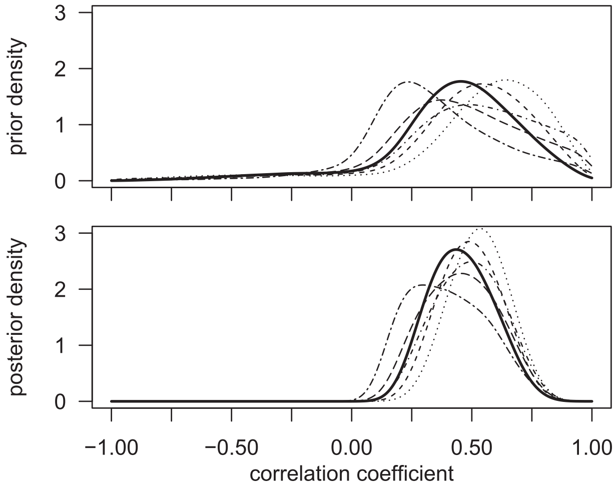

FIG. D1. (top) Different prior distributions on the correlation coefficient yield and (bottom) different posterior distributions (with corresponding line types). The thick solid lines correspond to the specifications of section 3c.

shown in Fig. D1. As expected, because of the small sample size, the posterior distributions vary considerably as a result of the variability of the prior. But the differences between the different prior distributions are 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 approximately 0.7 are shrunk toward a mode at approximately 0.5, which is smaller than the sample correlation of 0.62 for the data.

In section 3e, we have shown that there is a high posterior probability of an anomalously low signal-to-noise ratio of the model: $\Pr(\mathrm{SNR}_{\mathrm{obs}} > \mathrm{SNR}_{\mathrm{mod}}) > 0.99$. This probability is sensitive to the choice of the prior parameters. Note that, changing only the prior distribution of σ_s^2, the prior distribution of the correlation changes but the prior probability $\Pr(\mathrm{SNR}_{\mathrm{obs}} > \mathrm{SNR}_{\mathrm{mod}}) \approx 0.5$ does not change. We found that, changing the prior of σ_s^2 in such a way that the correlation prior becomes more pessimistic, the posterior probability for $(\mathrm{SNR}_{\mathrm{obs}} > \mathrm{SNR}_{\mathrm{mod}})$ decreases. For the prior specifications that yield the most pessimistic prior expectation of the correlation of approximately 0.3 in Fig. D1, the posterior probability of an anomalous SNR reduces to approximately 0.85.

REFERENCES

Annan, J., and J. Hargreaves, 2010: Reliability of the CMIP3 ensemble. *Geophys. Res. Lett.*, **37**, L02703, doi:10.1029/2009GL041994.

Bartholomew, D. J., M. Knott, and I. Moustaki, 2011: *Latent Variable Models and Factor Analysis: A Unified Approach.* Wiley Series in Probability and Statistics, Vol. 899, John Wiley & Sons, 294 pp.

Bradley, A. A., S. S. Schwartz, and T. Hashino, 2004: Distributions-oriented verification of ensemble streamflow predictions. *J. Hydrometeor.*, **5**, 532–545, doi:10.1175/1525-7541(2004)005<0532:DVOESP>2.0.CO;2.

Brooks, S., A. Gelman, G. Jones, and X.-L. Meng, 2011: *Handbook of Markov Chain Monte Carlo.* CRC Press, 611 pp.

Buonaccorsi, J. P., 2010: *Measurement Error: Models, Methods, and Applications.* CRC Press, 464 pp.

Chandler, R. E., 2013: Exploiting strength, discounting weakness: Combining information from multiple climate simulators. *Philos. Trans. Roy. Soc.*, **A371**, 20120388, doi:10.1098/rsta.2012.0388.

Doblas-Reyes, F., V. Pavan, and D. Stephenson, 2003: The skill of multi-model seasonal forecasts of the wintertime North Atlantic Oscillation. *Climate Dyn.*, **21**, 501–514, doi:10.1007/s00382-003-0350-4.

Eade, R., D. Smith, A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, and N. Robinson, 2014: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.*, **41**, 5620–5628, doi:10.1002/2014GL061146.

Efron, B., and R. J. Tibshirani, 1994: *An Introduction to the Bootstrap. Monogr. Stat. Appl. Probab.*, No. 57, CRC Press, 456 pp.

Everitt, B. S., 1984: *An Introduction to Latent Variable Models. Monogr. Stat. Appl. Probab.*, No. 7, Springer, 108 pp.

Feddersen, H., A. Navarra, and M. N. Ward, 1999: Reduction of model systematic error by statistical correction for dynamical seasonal predictions. *J. Climate*, **12**, 1974–1989, doi:10.1175/1520-0442(1999)012<1974:ROMSEB>2.0.CO;2.

Fuller, W. A., 1987: *Measurement Error Models.* John Wiley & Sons, 440 pp.

Gelman, A., and D. B. Rubin, 1992: Inference from iterative simulation using multiple sequences. *Stat. Sci.*, **7**, 457–472, doi:10.1214/ss/1177011136.

——, and C. P. Robert, 2013: "Not only defended but also applied": The perceived absurdity of Bayesian inference. *Amer. Stat.*, **67**, 1–5, doi:10.1080/00031305.2013.760987.

——, J. B. Carlin, H. S. Stern, and D. B. Rubin, 2004: *Bayesian Data Analysis.* Chapman and Hall/CRC, 690 pp.

Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, doi:10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, doi:10.1175/MWR2904.1.

Goddard, L., and Coauthors, 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dyn.*, **40**, 245–272, doi:10.1007/s00382-012-1481-2.

Hastie, T., R. Tibshirani, and J. Friedman, 2009: *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer Series in Statistics, Vol. 2, Springer, 745 pp., doi:10.1007/978-0-387-84858-7.

Jaynes, E. T., 2003: *Probability Theory: The Logic of Science.* Cambridge University Press, 753 pp.

Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* John Wiley & Sons, 292 pp.

Kang, D., and Coauthors, 2014: Prediction of the Arctic Oscillation in boreal winter by dynamical seasonal forecasting systems. *Geophys. Res. Lett.*, **41**, 3577–3585, doi:10.1002/2014GL060011.

Kharin, V. V., and F. W. Zwiers, 2003: Improved seasonal probability forecasts. *J. Climate*, **16**, 1684–1701, doi:10.1175/1520-0442(2003)016<1684:ISPF>2.0.CO;2.

Kumar, A., P. Peng, and M. Chen, 2014: Is there a relationship between potential and actual skill? *Mon. Wea. Rev.*, **142**, 2220–2227, doi:10.1175/MWR-D-13-00287.1.

Lindley, D. V., 2006: *Understanding Uncertainty.* John Wiley & Sons, 272 pp.

MacLachlan, C., and Coauthors, 2014: Global Seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system. *Quart. J. Roy. Meteor. Soc.*, **141**, 1072–1084, doi:10.1002/qj.2396.

Madden, R. A., 1976: Estimates of the natural variability of time-averaged sea-level pressure. *Mon. Wea. Rev.*, **104**, 942–952, doi:10.1175/1520-0493(1976)104<0942:EOTNVO>2.0.CO;2.

Mardia, K. V., J. T. Kent, and J. M. Bibby, 1979: *Multivariate Analysis.* Academic Press, 521 pp.

Moran, P., 1971: Estimating structural and functional relationships. *J. Multivar. Anal.*, **1**, 232–255, doi:10.1016/0047-259X(71)90013-3.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338, doi:10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.

——, and D. S. Wilks, 1998: A case study of the use of statistical models in forecast verification: Precipitation probability forecasts. *Wea. Forecasting*, **13**, 795–810, doi:10.1175/1520-0434(1998)013<0795:ACSOTU>2.0.CO;2.

Murphy, J. M., 1990: Assessment of the practical utility of extended range ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **116**, 89–125, doi:10.1002/qj.49711649105.

NCAR staff, Eds., 2014a: The Climate Data Guide: Hurrell North Atlantic Oscillation (NAO) Index (station-based), accessed 18 November 2014. [Available online at https://climatedataguide.ucar.edu/climate-data/hurrell-north-atlantic-oscillation-nao-index-station-based.]

——, Eds., 2014b: The Climate Data Guide: Hurrell North Atlantic Oscillation (NAO) Index (PC-based), accessed 18 November 2014. [Available online at https://climatedataguide.ucar.edu/climate-data/hurrell-north-atlantic-oscillation-nao-index-pc-based.]

Otto, F. E., C. Ferro, T. Fricker, and E. Suckling, 2012: On judging the credibility of climate projections. *Climatic Change*, **132**, 47–60, doi:10.1007/s10584-013-0813-5.

Pearl, J., 2000: *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 400 pp.

Riddle, E. E., A. H. Butler, J. C. Furtado, J. L. Cohen, and A. Kumar, 2013: CFSv2 ensemble prediction of the wintertime Arctic Oscillation. *Climate Dyn.*, **41**, 1099–1116, doi:10.1007/s00382-013-1850-5.

Robert, C. P., 2007: *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation (Springer Texts in Statistics).* Springer, 606 pp., doi:10.1007/0-387-71599-1.

Rougier, J., M. Goldstein, and L. House, 2013: Second-order exchangeability analysis for multimodel ensembles. *J. Amer. Stat. Assoc.*, **108**, 852–863, doi:10.1080/01621459.2013.802963.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660, doi:10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

Sansom, P. G., D. B. Stephenson, C. A. Ferro, G. Zappa, and L. Shaffrey, 2013: Simple uncertainty frameworks for selecting weighting schemes and interpreting multimodel ensemble climate change experiments. *J. Climate*, **26**, 4017–4037, doi:10.1175/JCLI-D-12-00462.1.

Scaife, A., and Coauthors, 2014: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, 2514–2519, doi:10.1002/2014GL059637.

Shi, W., N. Schaller, D. MacLeod, T. Palmer, and A. Weisheimer, 2015: Impact of hindcast length on estimates of seasonal climate predictability. *Geophys. Res. Lett.*, **42**, 1554–1559, doi:10.1002/2014GL062829.

Smith, D. M., A. A. Scaife, R. Eade, and J. R. Knight, 2014: Seasonal to decadal prediction of the winter North Atlantic Oscillation: Emerging capability and future prospects. *Quart. J. Roy. Meteor. Soc.*, doi:10.1002/qj.2479, in press.

Stan Development Team, 2014a: RStan: The R interface to Stan, version 2.5.0. Stan Development Team. [Available online at http://mc-stan.org/interfaces/rstan.html.]

——, 2014b: Stan modeling language user's guide and reference manual, version 2.5.0. Stan Development Team Tech. Rep., 408 pp. [Available online at http://mc-stan.org/.]

Stephenson, D. B., M. Collins, J. C. Rougier, and R. E. Chandler, 2012: Statistical problems in the probabilistic prediction of climate change. *Environmetrics*, **23**, 364–372, doi:10.1002/env.2153.

Tebaldi, C., R. L. Smith, D. Nychka, and L. O. Mearns, 2005: Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *J. Climate*, **18**, 1524–1540, doi:10.1175/JCLI3363.1.

Tippett, M. K., A. G. Barnston, and A. W. Robertson, 2007: Estimation of seasonal precipitation tercile-based categorical probabilities from ensembles. *J. Climate*, **20**, 2210–2228, doi:10.1175/JCLI4108.1.

Von Storch, H., and F. W. Zwiers, 2001: *Statistical Analysis in Climate Research.* Cambridge University Press, 496 pp.

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc.*, **131**, 965–986, doi:10.1256/qj.04.120.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2009: Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? *Mon. Wea. Rev.*, **137**, 1460–1479, doi:10.1175/2008MWR2773.1.