Parameter uncertainty in forecast recalibration

Stefan Siegert*, Philip G. Sansom and Robin Williams

Exeter Climate Systems, University of Exeter, Laver Building, North Park Road, Exeter, EX4 4QE, UK

November 24, 2015

Abstract

Ensemble forecasts of weather and climate are subject to systematic biases in the ensemble mean and variance, leading to inaccurate estimates of the forecast mean and variance. To address these biases, ensemble forecasts are post-processed using statistical recalibration frameworks. These frameworks often specify parametric probability distributions for the verifying observations. A common choice is the Normal distribution with mean and variance specified by linear functions of the ensemble mean and variance. The parameters of the recalibration framework are estimated from historical archives of forecasts and verifying observations. Often there are relatively few forecasts and observations available for parameter estimation, and so the fitted parameters are also subject to uncertainty. This artefact is usually ignored. This study reviews analytic results that account for parameter uncertainty in the widely used Model Output Statistics recalibration framework. The predictive bootstrap is used to approximate the parameter uncertainty by resampling in more general frameworks such as Non-homogeneous Gaussian Regression. Forecasts on daily, seasonal and annual time scales are used to demonstrate that accounting for parameter uncertainty in the recalibrated predictive distributions leads to probability forecasts that are more skilful and reliable than those in which parameter uncertainty is ignored. The improvements are attributed to more reliable tail probabilities of the recalibrated forecast distributions.

1 Introduction

Raw forecasts produced by numerical atmosphere-ocean models are often not representative of the real world. Modellers have long realised that there are systematic discrepancies between model simulations and the real world. Glahn and Lowry [1972] suggested that a linear transformation should be applied to weather model forecasts to issue predictions of real world observables. Similarly, Leith [1974]

*Corresponding author address: s.siegert@exeter.ac.uk

suggested that "a final regression step is needed" to get the "best estimate of the true state".

The past 50 years has seen a shift in focus from forecasts that are deterministic in nature to probability forecasts that represent the forecaster's uncertainty of the future, while also providing a point forecast. To this end, ensemble forecasts have become widely used in the fields of climate science and numerical weather prediction. An ensemble forecast is simply a collection of forecasts that differ in one or more of the model physics, resolution, or initial conditions.

Statistical adjustment to better fit numerical model output to real world observations is also known as forecast recalibration. A parametric statistical framework is specified that takes the raw numerical model forecast as an input, and outputs an estimate of the real world. To avoid confusion due to overuse of the word "model", the term "statistical framework" is used in place of the more common "statistical model". One of the simplest statistical forecast adjustments is the removal of a constant bias. The underlying assumption is that the observation is equal to the model output plus a constant offset. The offset is represented by an unknown parameter that must be estimated from training data.

A variety of more flexible recalibration frameworks have been proposed. In general, the choice of the recalibration framework depends on the forecast quantity, and on the assumptions about the forecast errors; for example, temperature forecasts require different recalibration frameworks than precipitation forecasts. Two of the most well-known frameworks are Model Output Statistics [MOS, Glahn and Lowry, 1972] and Non-homogeneous Gaussian Regression [NGR, Gneiting et al., 2005]. MOS is equivalent to Normal linear regression of the observations on the model output. NGR extends MOS by allowing the predictive uncertainty to depend on the ensemble spread. The more flexible the framework, the greater the number of parameters to be estimated.

It is common practice to recalibrate the forecast using the estimated parameters as the "correct parameter values", i.e., by treating them as known constants. However, the parameters are estimated from a finite sample of training data and so are also subject to uncertainty. By naïvely using the fitted parameters in issuing probability forecasts, the uncertainty in the parameter estimates is ignored.

The problem of accounting for parameter uncertainty when recalibrating climate and weather forecasts is not unknown. Glahn et al. [2009b] state that the predictive distribution of forecasts recalibrated by linear regression should be a Student's t-distribution with inflated variance. Related remarks can be found in Mason and Mimmack [2002], Bröcker and Smith [2008], and Unger et al. [2009]. Forecast recalibration using dynamic linear models (Kalman filtering) also leads to forecasts that follow a t-distribution [Sohn et al., 2003, Pagowski et al., 2006]. Bayesian methods can also be used to account for parameter uncertainty in forecasts recalibrated by linear regression [Marty et al., 2014], or by latent variable methods [Siegert et al., 2015].

This study investigates the effect of accounting for parameter uncertainty on the reliability and skill of recalibrated probability forecasts. Section 2 describes

analytic methods to account for parameter uncertainty in MOS, and proposes a simple bootstrap method to account for parameter uncertainty in NGR. Section 3 applies the methods developed in Section 2 to three meteorological forecast data sets: an annual mean temperature forecast, a seasonal forecast of the North-Atlantic Oscillation, and a short-range 48-hour temperature forecast. All three data sets demonstrate that accounting for parameter uncertainty improves both the skill and reliability of recalibrated forecasts. Section 4 concludes with a summary and discussion.

2 Methodology

2.1 Model Output Statistics (MOS)

The original application of MOS was statistical downscaling, i.e., to produce forecasts of quantities that are not explicitly modelled numerically, such as surface winds or probability of precipitation [Glahn and Lowry, 1972]. However, MOS has been widely used for the statistical recalibration of both deterministic and ensemble forecasts [Kharin and Zwiers, 2003, Tippett et al., 2005, Glahn et al., 2009b]. As noted above, MOS is equivalent to Normal linear regression. The future (unknown) observation is represented by a linear function of the numerical model output, with Normally distributed forecast errors. Let y_t denote the observed conditions at time t, and let m_t be the corresponding ensemble forecast mean. Then the MOS recalibration framework is given by

$$y_t = a + bm_t + c\varepsilon_t, \tag{1}$$

where ε_t is a standard Normally-distributed random variable, i.e., $\varepsilon_t \sim \mathcal{N}(0,1)$. For simplicity, we focus on simple linear regression. However, all the results presented in this paper extend easily to multiple linear regression, where the forecast mean depends on more than one input [e.g., Glahn et al., 2009a].

The parameters a, b, and c can be estimated by maximising the log-likelihood under the assumption that the errors $\varepsilon_t, t = 1, 2, \ldots, n$ are independent and follow a standard Normal distribution. Given a training set of n ensemble forecast means $m_1, \ldots, m_n$ and corresponding verifying observations $y_1, \ldots, y_n$, the unbiased maximum-likelihood estimates of a, b, and c^2 are

$$\hat{a} = \bar{y} - \frac{s_{my}}{s_m^2}\bar{m}, \tag{2a}$$

$$\hat{b} = \frac{s_{my}}{s_m^2}, \tag{2b}$$

$$\hat{c}^2 = \frac{1}{n-2}\sum_{t=1}^{n}\left(y_t - \hat{a} - \hat{b}m_t\right)^2, \tag{2c}$$

where $\bar{m}$ and $\bar{y}$ denote the overall sample means of the ensemble means $m_1, \ldots, m_n$ and observations $y_1, \ldots, y_n$ in the training sample, and s_{my} and s_m^2 denote the sample covariance between ensemble means and observations and the sample

variance of the ensemble means, respectively. Note the division by $n-2$ in Eqn. 2c, to account for the fact that two mean parameters (a and b) have been estimated [Draper et al., 1998, Chp. 1]. The fitted parameters $\hat{a}$, $\hat{b}$ and $\hat{c}$ can then be used to recalibrate new forecasts for yet unknown observations.

2.2 Non-homogeneous Gaussian Regression

Recalibration using MOS explicitly assumes that the predictive variance is constant for all forecasts, i.e., equal to c^2. In practice, the forecast uncertainty might be different on different occasions due to varying error growth rates, and more or less predictable weather regimes. If the numerical model can reproduce this variability, then there might be useful information not only in the ensemble mean, but also in the ensemble variance. Recalibration using NGR aims to exploit systematic relationships between the ensemble variance and the variance of forecast errors [Gneiting et al., 2005, NGR]. The NGR forecast mean is a linear function of the ensemble mean, m_t, identical to MOS. But unlike MOS, the forecast variance at time t is a linear function of the ensemble variance, v_t. The NGR recalibration framework for the observation y_t is thus

$$y_t \sim \mathcal{N}\left(a + bm_t, c + dv_t\right). \tag{3}$$

Note that NGR recalibration with the parameter d fixed at zero is equivalent to MOS recalibration.

Gneiting et al. [2005] suggest parameter estimation for NGR by numerical minimisation of the Continuous Ranked Probability Score (CRPS). However, Williams et al. [2014] found little advantage of minimum-CRPS estimation compared to maximum-likelihood estimation. To maintain continuity with the MOS parameter estimators, the NGR recalibration parameters are estimated by maximising the log-likelihood function

$$\ell_{\mathrm{NGR}} \propto -\sum_{t=1}^{n}\left[\log\left(c + dv_t\right) + \frac{\left(y_t - a - bm_t\right)^2}{c + dv_t}\right]. \tag{4}$$

Since closed form solutions are not available for the maximum-likelihood estimates of the NGR parameters, numerical optimisation is used.[1] To ensure that the forecast variance is positive, the variance scale parameter is represented by $d = \delta^2$ and optimised over δ [Gneiting et al., 2005].

2.3 Parameter uncertainty in MOS: Analytic results

Suppose the maximum-likelihood estimates $\hat{a}$, $\hat{b}$, and $\hat{c}$ from Equations 2a – 2c are known. These parameter estimates can be used to recalibrate a new ensemble mean forecast m^* for a yet unknown future observation y^*. It is

[1] Numerical optimisation is performed using the function `optim` in the `stats` package of the `R` statistical programming environment [R Core Team, 2015, version 3.2.0]

common practice to directly substitute the maximum-likelihood estimates into the MOS recalibration framework, so that the forecast distribution for y^* is

$$y^* \sim \mathcal{N}\left(\hat{a} + \hat{b}m^*, \hat{c}^2\right). \tag{5}$$

This recalibration framework has been used for probabilistic forecasting by, e.g., Kharin and Zwiers [2003], Tippett et al. [2005].

Results from linear regression [Draper et al., 1998, Chp. 1] show that the predictive distribution should also include parameter uncertainty. That is, the predictive distribution should not only include the estimated variance of the forecast errors $\hat{c}^2$, but should also account for the estimation uncertainty in the parameter estimates $\hat{a}$, $\hat{b}$ and $\hat{c}^2$. The sampling variances of the parameter estimates $\hat{a}$ and $\hat{b}$ lead to an increase of the predictive variance of the Normal distribution. Uncertainty due to estimation of the error variance c^2 leads to a transformation of the Normal distribution into a Student's t-distribution.

When parameter uncertainty is taken into account, the predictive distribution in linear regression becomes a t-distribution with inflated variance:

$$y^* \sim t_{n-2}\left(\hat{a} + \hat{b}m^*, \hat{c}^2\left[1 + \frac{1}{n} + \frac{(m^* - \bar{m})^2}{\sum_{t=1}^{n}(m_t - \bar{m})^2}\right]\right). \tag{6}$$

The forecast variance is inflated by a term that depends on both the sample size n, and the distance of the ensemble forecast mean m^* from the overall mean of the training forecasts, $\bar{m}$. The function $t_\nu(\mu, \sigma^2)$ denotes the non-standardized Student's t-distribution with ν degrees of freedom, location μ and scale σ (see appendix). In the limit as $\nu \to \infty$, the t-distribution converges to a Normal distribution. However, for small ν, the tails of the t-distribution are heavier than those of the Normal distribution. Therefore, the variances of the predictive distributions given by Equations 5 and 6 differ when the sample size n is small, or when $|m^* - \bar{m}|$ is large. Note that the forecast mean does not change.

The difference due to the adjustment for parameter uncertainty is illustrated in Figure 1. The standard Normal distribution $\mathcal{N}(0, 1)$ is compared to the non-standardized t-distribution $t_{18}(0, 1.1)$. The 18 degrees of freedom correspond to a sample size of $n = 20$, and a 10% inflated variance corresponds to a forecast m^* that differs from the overall forecast mean of the training data by one standard deviation. The example can thus be considered a typical case for what one might expect to see with a training sample of climate forecasts. Figure 1 shows that the difference due to the adjustment for parameter uncertainty is generally small, and so one should not expect the differences in forecast quality to be substantial.

2.4 Parameter uncertainty in NGR: The predictive bootstrap

For NGR recalibration, Gneiting et al. [2005] suggest substituting the parameter estimates $\hat{a}, \hat{b}, \hat{c}$ and $\hat{d}$, as well as the ensemble mean m^* and variance v^*, directly

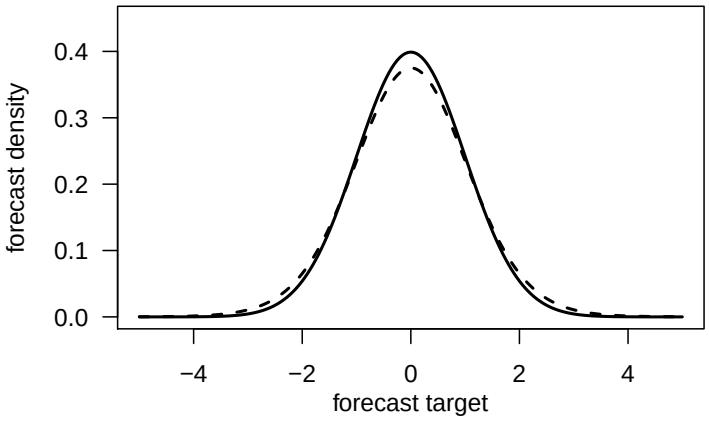

Figure 1: Illustration of difference between the standard Normal distribution $\mathcal{N}(0,1)$ (solid line) and the non-standardized t-distribution $t_{18}(0,1.1)$ (dashed line).

into Equation 3. The forecast distribution for the future observation y^* is then given by

$$y^* \sim \mathcal{N}\left(\hat{a} + \hat{b}m^*, \hat{c} + \hat{d}v^*\right). \tag{7}$$

This approach has also been used to recalibrate probability forecasts in a number of other studies, e.g., Hagedorn et al. [2008], Kann et al. [2009].

However, in keeping with the discussion in Section 2.3, simply substituting the parameter estimates and issuing forecasts with a Normal distribution ignores parameter uncertainty. No analytic expression for the NGR forecast distribution that accounts for parameter uncertainty has been published, to date. Therefore, a means of approximating the parameter uncertainty, and accounting for the parameter uncertainty in the predictive distribution is required. The parameter uncertainty can be estimated non-parametrically by bootstrapping [Efron, 1982]. Generating predictive distributions using bootstrap resampling is known as predictive bootstrapping, and was originally proposed by Harris [1989]. Given a historical archive of ensemble mean forecasts m_t, ensemble variances v_t, and corresponding observations y_t, for times $t = 1, 2, \ldots, n$, the following bootstrap resampling protocol is proposed:

1. Generate a new training data set of size n by randomly sampling n times *with replacement* from the available pairs of historical forecasts and observations;

2. Compute the maximum likelihood NGR parameter estimates using this new training data set; and

3. Repeat steps 1. and 2. K times.

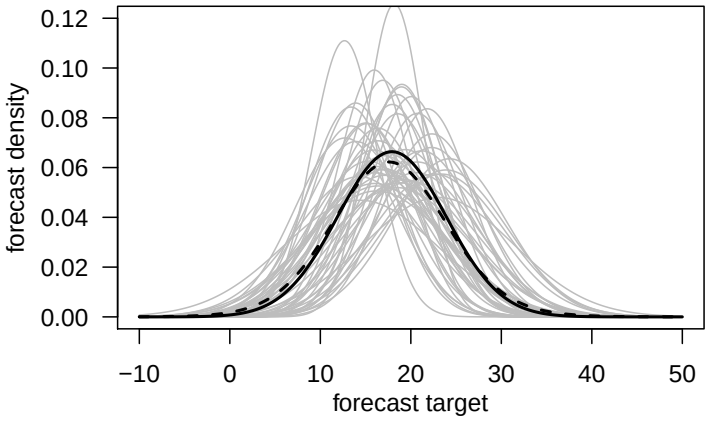

Figure 2: Illustration of the predictive bootstrap to account for parameter uncertainty in NGR using a training data set of $n = 19$. The Normal forecast distribution using maximum likelihood estimates of the NGR parameters (solid black line), 50 Normal distributions with maximum likelihood parameter estimates obtained by resampling the training data set (solid gray lines), and the distribution obtained by the averaging the 50 bootstrap distributions (dashed black line).

The kth resampling leads to the bootstrap parameter estimates $\tilde{a}_k$, $\tilde{b}_k$, $\tilde{c}_k$, and $\tilde{d}_k$. The collection of K bootstrap parameter estimates approximates the parameter uncertainty distribution.

The objective is to generate a recalibrated forecast for the unknown value y^* of the observation, using the ensemble mean forecast m^* and ensemble variance v^*. Each set of bootstrapped parameter estimates $(\tilde{a}_k, \tilde{b}_k, \tilde{c}_k, \tilde{d}_k), k = 1, 2, \ldots, K$ leads to a Normally-distributed forecast given by Equation 7. The K bootstrap samples are combined into a single predictive distribution by calculating the equally weighted average over the individual Normal distributions, thereby producing a Normal mixture distribution (see appendix). The bootstrap forecast distribution itself is therefore not a Normal distribution.

The predictive bootstrap is illustrated in Figure 2. The effect of averaging over the bootstrapped Normal distribution is similar to the adjustment for parameter uncertainty in MOS: The variance of the distribution is increased, and the tails are made heavier. Unlike MOS, the mode of the predictive distribution can be different after accounting for parameter uncertainty. By exploring different values for the estimated slope parameter $\hat{b}$, bootstrapping also inflates the forecast variance when the forecast mean m^* is far from the mean of the training data $\bar{m}$.

The assumptions underlying the predictive bootstrap are quite different from those of the analytic method in Section 2.3. Both approaches assume that the

training data represent independent and identically distributed samples (conditional on the parameters) from some unknown distribution. However, the bootstrap approximates the distribution of the parameters, without making prior assumptions about the form of the distribution, e.g., Normal, Student t, etc.. Therefore, bootstrapping is useful for frameworks such as NGR, where parametric inference is difficult or impossible.

2.5 Forecast verification: CRPS, Ignorance, PIT histogram

This section outlines the forecast verification measures used in this paper to assess the quality and improvements of probabilistic forecasts. Only the general forms of the verification measures are provided here. Equations for specific distributions can be found in the appendix.

The Ignorance score is a verification score to evaluate probability density forecasts [Roulston and Smith, 2002]. If the forecast probability density function (pdf) is $f(x)$, and the verifying observation is y, then the Ignorance score is given by

$$ign(f, y) = -\log_2 f(y), \tag{8}$$

i.e., the negative logarithm of the forecast density evaluated at the observation. The Ignorance score is a local score, i.e. it only depends on the value of the forecast assigned to the verifying observation. If the basis 2 is used, Ignorance differences are measured in bits. An Ignorance difference of $\Delta > 0$ bits between forecast A and forecast B implies that forecast B has assigned 2^{Δ} times more probability density to the verifying observation than forecast A. Forecast B, having lower Ignorance score than forecast A, can thus be considered to be the "better" forecast. Time-averaged Ignorance differences are used as summary measures of relative forecast performance.

The Continuous Ranked Probability Score (CRPS) is designed to evaluate a cumulative forecast distribution (cdf) $F(x)$ for a scalar observation y [Matheson and Winkler, 1976, Hersbach, 2000]. The general form of the CRPS is

$$crps(F, y) = \int_{-\infty}^{+\infty} \left[F(x) - H(x - y)\right]^2 dx, \tag{9}$$

where $H(x)$ is the Heaviside step-function, i.e., $H(x) = 0$ for $x \leq 0$ and $H(x) = 1$ otherwise. Unlike the Ignorance score, the CRPS is a non-local score. The CRPS depends not only on the value assigned to the observation, but also on how much forecast probability is concentrated near the observation, i.e., the CRPS is sensitive to the distance of the bulk of the forecast distribution from the observation. For a deterministic forecast (i.e. where $F(x)$ is itself a step function), the CRPS is equal to the absolute difference between forecast and observation, and the CRPS vanishes for a perfect deterministic forecast where $F(x) = H(x-y)$. Therefore, the CRPS can be interpreted as a distance measure between forecast and observation, and a lower CRPS value can be taken as an indication of a "better" forecast. Note that the notion of "better" depends on

the score, e.g., forecast A can perform "better" than forecast B in terms of the Ignorance score, but "worse" in terms of the CRPS.

The CRPS is measured on the scale of the forecast target x. The Continuous Ranked Probability Skill Score (CRPSS) provides a dimensionless measure of the difference in forecast performance. If the time-averaged CRPS of forecasts A and B are $CRPS_A$ and $CRPS_B$, the relative improvement of forecast B over forecast A is given by

$$CRPSS = \frac{CRPS_A - CRPS_B}{CRPS_A}. \tag{10}$$

A positive (negative) CRPSS indicates an improvement (deterioration) of forecast B compared to forecast A. CRPSS close to zero indicates that the forecasts are equally good. The CRPSS is bounded above at unity for a perfect forecast B, but has no lower bound.

Probabilistic forecasts should be issued such that the verifying observations behave like random draws from the forecast distributions. Such forecasts are referred to as being reliable or well-calibrated [Gneiting et al., 2007]. For reliable forecasts, the observations fall on average equally often between the 0 and 5 percentiles, between the 5 and 10 percentiles, etc., of the forecast distribution. Therefore, counting how often each percentile interval is occupied by the observation provides a simple test of forecast reliability. In practice, the probability integral transform (PIT) of each forecast pdf $f_t(x)$ is calculated, which is the mass of forecast probability below its verifying observation:

$$pit_t := PIT(f_t, y_t) = \int_{-\infty}^{y_t} f_t(x)dx. \tag{11}$$

For example, if the observation y_t falls between the 5th and 10th percentiles of the forecast density $f_t(x)$, then the value of pit_t will be between 0.05 and 0.10. Since each percentile interval should be equally likely on average for a reliable forecast, reliability can be checked by observing the shape of the PIT histogram. Reliable forecasts have a flat PIT histogram, and a non-flat PIT histogram indicates unreliable forecasts. In general, $\cup$-shaped histograms indicate underdispersed forecast distributions, i.e., overconfident forecasts. Conversely, $\cap$-shaped histograms suggest overdispersed forecast distributions (underconfident forecasts), while sloping histograms are indicative of a systematic bias of the forecast mean [Hamill, 2001].

3 Results

3.1 CanCM4 gridded annual temperature forecasts

The effect of accounting for parameter uncertainty was evaluated in forecasts of near-surface (2m) temperature generated by the fourth version of the Canadian coupled ocean-atmosphere general circulation model [CanCM4; Merryfield et al., 2011]. Forecast ensembles of 10 initial condition members were initialized on

1 January every year between 1960 and 2010. The forecast target was the mean temperature averaged over the first 12 months after initialization. Verifying observations were taken from the HadCRUT3v data set [Brohan et al., 2006]. Only grid boxes with a complete observational record were included in the analysis. The conclusions were not found to be sensitive to the choice of verification data; comparable results were obtained when verifying against the ERA-Interim reanalysis [Dee et al., 2011] and all grid boxes. MOS-recalibrated probability forecasts were computed for each year in the period 1991–2010. It was found that the ensemble variance was not a skilful predictor of forecast errors, and so there was no advantage to using NGR recalibration. Forecasts were recalibrated using only information available up to the time of the forecast and evaluated out-of-sample. The recalibration parameters for each forecast were estimated using the previous 25 years of forecasts and observations, after linearly detrending both data sets.

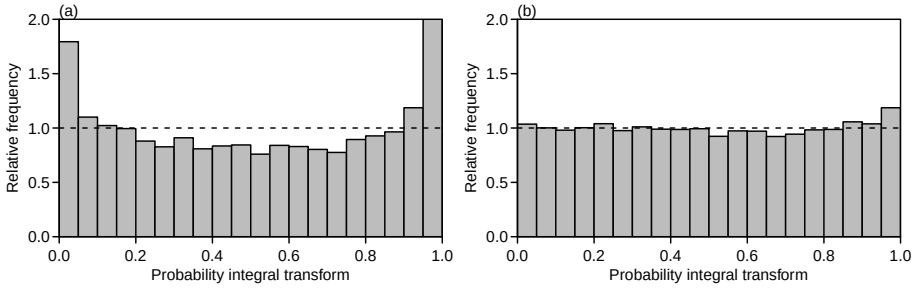

Figure 3: PIT histograms of recalibrated CanCM4 forecasts (a) before, and (b) after accounting for parameter uncertainty. Before accounting for parameter uncertainty, the observations fall into the tails of the forecast distribution too often.

The PIT histogram of the recalibrated forecasts without parameter uncertainty (Equation 5) appear ∪-shaped, indicating underdispersed probability forecasts (Figure 3a). As a result of the underdispersion, the 90% prediction intervals cover the verifying observations only 81% of the time. Therefore, the forecasts without parameter uncertainty are not well calibrated. In contrast, the PIT histogram of the recalibrated forecasts after accounting for parameter uncertainty (Equation 6) is almost completely flat (Figure 3b). Each 5% interval of the predictive distributions covers the verifying observation roughly 5% of the time, and the 90% prediction intervals cover the verifying observations 89% of the time. The recalibrated forecasts including parameter uncertainty appear to be well calibrated.

The time-averaged difference in the Ignorance score between the recalibrated forecasts before and after accounting for parameter uncertainty is positive at most grid boxes (Figure 4a). Positive Ignorance differences indicate that the forecasts that account for parameter uncertainty assign more probability density

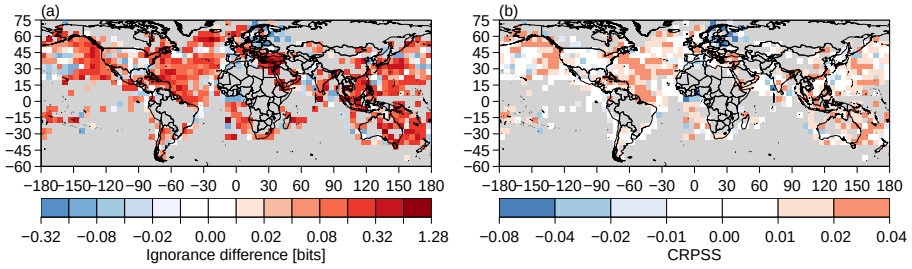

Figure 4: Improvement in overall skill of the recalibrated CanCM4 forecasts after accounting for parameter uncertainty, measured by (a) the difference in the Ignorance score, and (b) the CRPSS. Positive score differences / skill scores indicate improved average scores after accounting for parameter uncertainty.

to the observations than those that do not, i.e., are more skilful.

The CRPSS is also positive in the majority of grid boxes, supporting the conclusion that the forecasts that account for parameter uncertainty are on average more skilful (Figure 4b).

3.2 Met Office seasonal NAO forecasts

This section presents a case study on seasonal climate forecasts using a more complicated recalibration framework. The data consist of 20 years of historical seasonal ensemble forecasts of the North Atlantic Oscillation (Figure 5). The ensembles were produced by the UK Met Office Global Seasonal prediction system GloSea5 [MacLachlan et al., 2014]. The forecasts were initialised using a lagged initialisation around 1 November each year between 1992 and 2011. The forecast target is the average North Atlantic Oscillation (NAO) index between December and February (DJF), measured as a pressure difference between stations situated in the Azores and Iceland [Scaife et al., 2014].

The sample correlation coefficient between the ensemble means and the observations is 0.62. The correlation between the ensemble standard deviation and absolute error of the ensemble mean is also high at 0.45 (or 0.3 when the very influential year 2010 is excluded). Previous studies found that the skill of these forecasts can be improved by linear transformations of both the ensemble mean and spread [Eade et al., 2014]. Therefore, NGR recalibration was used to allow for linear adjustment not only of the predictive mean, but also the predictive variance. Due to the small sample size, forecasts were evaluated by leave-one-out cross-validation, i.e., each forecast was recalibrated using the other 19 as the training set. Due to the long time scale under consideration, each forecast occasion can reasonably be assumed to be independent, and so the leave-one-out approach is justifiable.

Figure 2 illustrates the predictive distributions for the year 1997, issued as a Normal distribution with the maximum likelihood NGR parameter estimates,

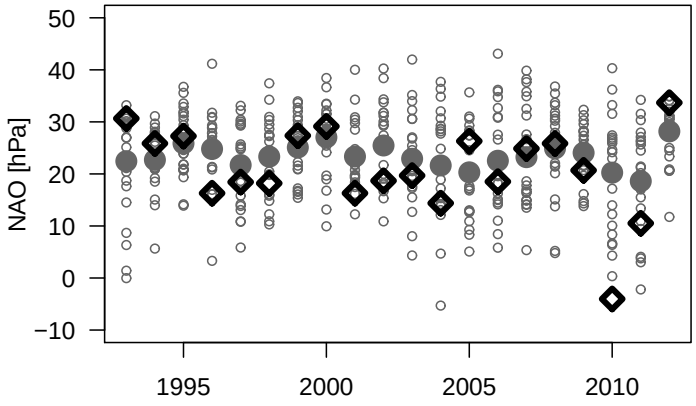

Figure 5: GloSea5 NAO ensemble forecasts (small circles), ensemble mean forecasts (large filled circles) and verifying observations (black diamonds) plotted over the verification year.

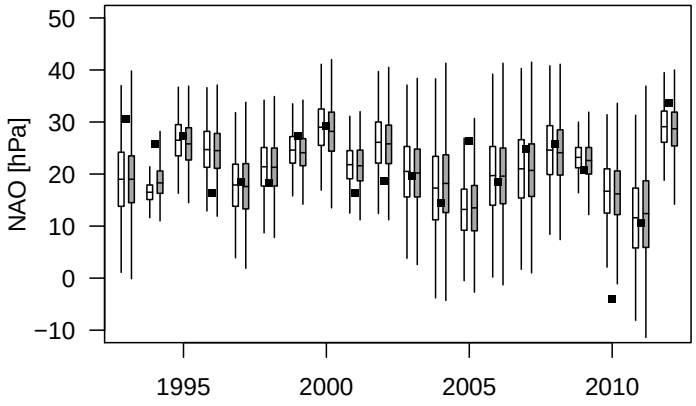

Figure 6: NAO observations (black markers) and their NGR predictive distributions using the ensemble data shown in Figure 5. Distributions without parameter uncertainty (white boxes) and with parameter uncertainty (gray boxes) are depicted by box-and-whiskers plots, where boxes indicate the inter-quartile range and median, and whiskers extend from the 1 to the 99 percentile.

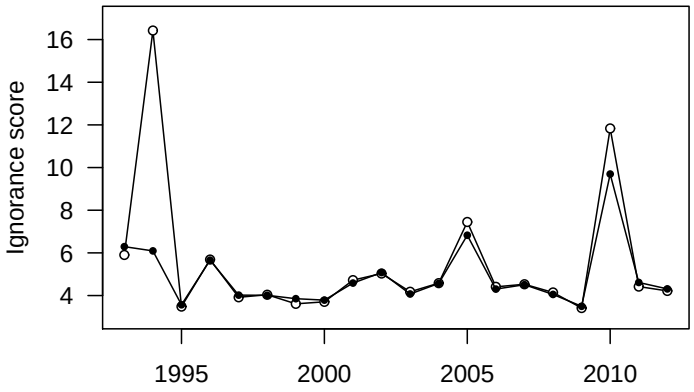

Figure 7: Ignorance scores of the forecast distributions shown in Figure 6, without accounting for parameter uncertainty (open circles) and after accounting for parameter uncertainty (full circles).

and issued as a mixture of Normal distributions based on 500 bootstrap replicates. The Normal distributions generated from the individual bootstrap resamples vary considerably, indicating large parameter uncertainty. The variance of the averaged bootstrap distribution is larger, and the tails are heavier than for the Normal distribution that uses only the maximum-likelihood parameters. Figure 6 shows the recalibrated predictive distributions for the years 1993-2012, with and without parameter uncertainty, and their verifying observations. The forecast distributions with parameter uncertainty have slightly heavier tails, and their medians are on average slightly closer to the climatological mean.

Figure 7 shows the Ignorance scores of the individual forecasts. Whenever the observation falls close to the bulk of the forecast distributions, the Ignorance scores for the two distributions are very similar. There are three cases where the observation falls well into the tail of the forecast distributions (1994, 2005 and 2010). For these tail events, the Ignorance score improves when parameter uncertainty is taken into account. The predictive bootstrap leads to heavier tails, and thus to higher probabilities being assigned to "unexpected events".

The results of the case study presented in this Section suggest that the predictive bootstrap increases the forecast skill of recalibrated GloSea5 NAO forecasts. However, only 20 data points are considered, which does not allow for a robust statistical analysis of the forecast skill. A larger dataset of numerical weather predictions is analysed in the next Section.

3.3 NCEP short range temperature forecasts

Daily forecasts of near-surface (2m) temperature with a 48 hour lead time were also analysed. The forecasts were taken from version 2 of the refore-

cast project, hosted by the National Oceanic and Atmospheric Administration, USA [Hamill et al., 2013]. The ensemble forecasts are approximately equivalent to those issued by the operational global ensemble forecasting system of the national centre for environmental prediction (NCEP). Ten member initial condition forecasts were issued at 00 UTC each day for a grid point close to New York City, USA (40N, 74W). The forecasts covered the period 26 May 1990 - 15 September 2014, giving a total of $n = 8,879$ forecasts and verifying observations. The analyses (i.e., the control forecast at 0 lead time) were used as verifying observations.

Preliminary investigations showed that NGR recalibration yielded more skilful forecasts than simple MOS recalibration. Therefore, the NGR recalibrated forecasts were used throughout. Recalibration parameters were estimated separately for all n forecasts, using data from a rolling training window of prespecified size w. That is, for each forecast, the w previous pairs of forecasts and observations were used as training data, such that all forecasts were evaluated out-of-sample. The rolling training window allows the recalibration scheme to adapt to non-stationarities in the hindcast data, such as the updating of the data assimilation scheme in 2011, or the dependency of the forecast bias on the time of the year [Hamill et al., 2013]. Parameter uncertainty was accounted for using the bootstrap approach described in Section 2.4. 50 bootstrap replicates were used for each forecast. Increasing the number of bootstrap replicates was found to provide only very small improvements in forecast skill, at considerable computational expense due to the large sample of forecasts to be evaluated.

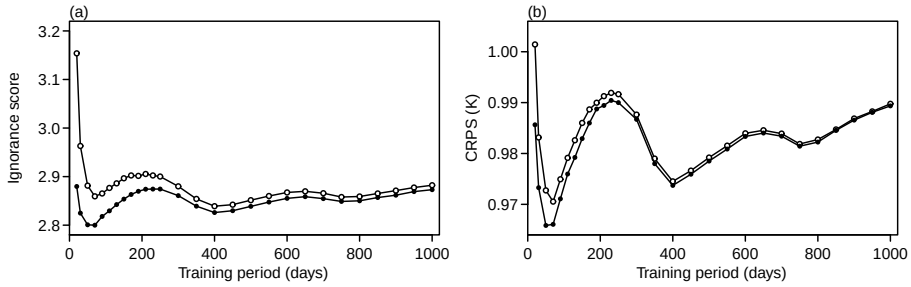

Figure 8: (a) Ignorance scores, and (b) CRPS as a function of training sample size for the recalibrated NCEP forecasts before (open circles) and after (filled circles) accounting for parameter uncertainty.

Figure 8a shows the Ignorance scores of the NGR-recalibrated forecasts as a function of the size of the rolling training window, before and after accounting for parameter uncertainty. The improvements in forecast skill produced by accounting for parameter uncertainty are evident for both small and large training samples. As expected, the scores converge for very large training datasets. It is encouraging that the predictive bootstrap yields improved forecasts even in relatively data-rich settings with hundreds of historical forecasts and observations,

where one might expect the effect of parameter uncertainty to be negligible. The CRPS values shown in Figure 8b are qualitatively similar to those of the Ignorance scores. The optimum training period after accounting for parameter uncertainty is 50 days for both scores. Before accounting for parameter uncertainty, the optimal training period for CRPS is similar (60 days). However, the optimal training period for the Ignorance score is 400 days without accounting for parameter uncertainty. For CRPS, the effect of the training length is large compared to the effect of parameter uncertainty, while the effects are of comparable magnitude for the Ignorance score.

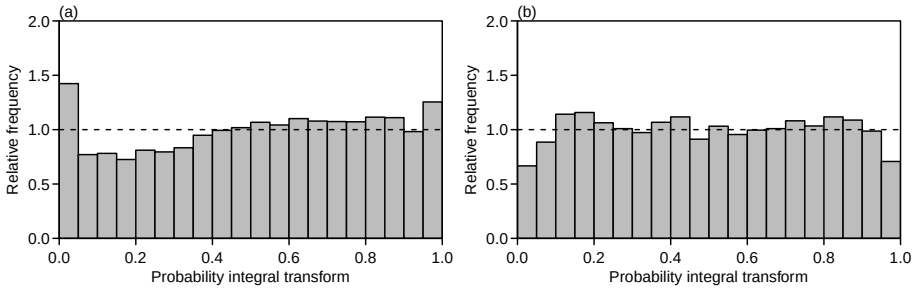

Figure 9: PIT histograms of recalibrated NCEP forecasts (a) before, and (b) after accounting for parameter uncertainty. Before accounting for parameter uncertainty, the forecasts show evidence of underdispersion and bias. After accounting for parameter uncertainty, the forecasts appear slightly overdispersed.

Figures 9a and b show PIT histograms after recalibration based on a rolling training period of 50 days, before and after accounting for parameter uncertainty. The overpopulation of the outer bins of the PIT histograms when parameter uncertainty was not accounted for is indicative of forecast distributions whose tails are too light. Accounting for parameter uncertainty by bootstrap resampling results in PIT histograms that are closer to being uniform. Figure 9b suggests that the resampling method has slightly overcompensated for the light tails, thus leading to overdispersive forecasts. In both cases, the PIT histograms of the bootstrap forecasts suggest remaining forecast biases, which might be an indication of a poor fit of the NGR recalibration framework to the data. Refinements of the underlying NGR framework are beyond the scope of this paper.

4 Discussion and conclusions

Parameter estimates in forecast recalibration frameworks are subject to uncertainty, particularly when estimated with small training samples. The effect of parameter uncertainty on the reliability and skill of probability forecasts has received little attention in the climate and meteorology literature. This study

has presented two methods of accounting for parameter uncertainty in recalibrated forecasts. Analytic results are available for MOS recalibration. Parameter uncertainty in more complex recalibration frameworks can be estimated by bootstrapping. The results presented here demonstrate that accounting for parameter uncertainty can improve the reliability and skill of recalibrated forecasts across a range of time scales.

The examples demonstrated here are representative of findings across a number of forecast models, variables and time scales. In some cases, accounting for parameter uncertainty does not improve forecast reliability and skill. For example, accounting for parameter uncertainty did not improve seasonal average European temperature forecasts from the ECMWF System4 [Molteni et al., 2011] at a lead time of 3 months. The PIT histogram did not change, and the verification scores improved at only 50% of grid boxes, leaving the average forecast skill unchanged. However, no cases were identified where accounting for parameter uncertainty leads to forecasts that are less reliable or less skilful on average.

The main effects of accounting for parameter uncertainty are the inflation of the forecast variance, and an increase in the weight of the tails of the forecast distribution. The wider, heavy tailed forecast distributions improve reliability by generating forecast distributions that are less underdispersive. This can be seen from the PIT histograms, which are less $\cup$-shaped after accounting for parameter uncertainty. Overall forecast skill is also improved, but the size of the improvement depends on the score. The Ignorance score is more sensitive to low-probability events than the CRPS. Since the main effect of accounting for parameter uncertainty is to improve the tails of the forecast distributions, the relative improvement in the Ignorance score is larger than that of the CRPS.

The effect of accounting for parameter uncertainty is largest for small training samples. The amount of training data can be limited by the available observational record, by strong temporal and spatial correlations in the data, or by the computational expense of generating long hindcast experiments. However, small training samples are often deliberately chosen even in data-rich situations such as weather forecasting. The use of a rolling training window allows the recalibration to adapt to changes in background conditions, or changes to the forecast model. Alternative methods might be considered so that all prior data are included in the training sample but with decreasing weight given to older data. Another possibility would be to use the idea of analogues, and only calibrate using prior data that are similar to conditions observed at the time the forecast is initialized.

Analytic expressions for the forecast uncertainty are possible for some more complex recalibration frameworks. The predictive bootstrap is easily applicable to almost any recalibration framework. Bootstrapping can be modified to account for temporal dependence between the training data by block resampling [Davison and Hinkley, 1997, Chp. 8]. Alternative methods of estimating the parameter uncertainty include parametric bootstrapping, and asymptotic approximations of the parameter uncertainty. Bayesian methods lead to similar analytic results in the case of MOS recalibration, and computational Bayesian

techniques can be used for more complex frameworks.

This study has demonstrated that accounting for parameter uncertainty in probability forecasts leads to measurable improvements in both reliability and skill. Other researchers and practitioners are encouraged to investigate and adopt the methods proposed here, and to develop alternative methods for more complex recalibration frameworks.

Acknowledgments

The authors would like to thank Adam Scaife and the members of the monthly-to-decadal prediction group at the Met Office Hadley Centre for providing the GloSea5 data. The authors would also like to thank Nathan Owen, Chris Ferro, David Stephenson, Tom Hamill, and an anonymous reviewer for helpful comments during the preparation of this manuscript. Stefan Siegert was supported by the European Union Programme FP7/2007-13 under grant agreement 3038378 (SPECS). Philip Sansom was supported by a grant from the National Oceanic and Atmospheric Administration (NOAA) NA12OAR4310086. The views expressed herein are those of the authors and do not necessarily reflect the views of their funding bodies or any of their subagencies.

Appendices

The non-standardized t-distribution

The pdf of the non-standardized t-distribution with location μ, scale σ and degrees-of-freedom ν is given by [West and Harrison, 1997]

$$p\left(x;\nu,\mu,\sigma^2\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}}\left[1+\frac{1}{\nu}\frac{(x-\mu)^2}{\sigma^2}\right]^{-\frac{\nu+1}{2}}. \tag{12}$$

Normal mixture distribution

Let $\varphi(x)$ and $\Phi(x)$ denote the pdf and cdf of the standard Normal distribution respectively. If the forecast pdf $f(x)$ is a mixture of K Normal distributions with weights $\omega_1,\cdots,\omega_K$ (non-negative and summing to one), means $\mu_1,\cdots,\mu_K$ and variances $\sigma_1^2,\cdots,\sigma_K^2$, then the forecast pdf $f(x)$ is given by

$$f(x) = \sum_{k=1}^{K}\omega_k\frac{1}{\sigma_k}\varphi\left(\frac{x-\mu_k}{\sigma_k}\right), \tag{13}$$

and the forecast cdf $F(x)$ is

$$F(x) = \sum_{k=1}^{K}\omega_k\Phi\left(\frac{x-\mu_k}{\sigma_k}\right). \tag{14}$$

The Ignorance Score

If the forecast pdf $f(x)$ is Normal with mean μ and variance σ^2, then the Ignorance score is given by

$$ign(f,y) = \left[\frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{(y-\mu)^2}{2\sigma^2}\right] / \log 2. \tag{15}$$

For a mixture of Normals, simply take the negative logarithm of Equation 13 after substituting appropriate values for the forecast means and variances and for the observation.

If the forecast pdf $f(x)$ has the form of a non-standardized t-distribution, then the Ignorance score is given by

$$ign(f,y) = \left[-\log\Gamma\left(\frac{\nu+1}{2}\right) + \log\Gamma\left(\frac{\nu}{2}\right) + \frac{1}{2}\log\left(\pi\nu\sigma^2\right) + \frac{\nu+1}{2}\log\left(1 + \frac{1}{\nu}\frac{(y-\mu)^2}{\sigma^2}\right)\right] / \log 2. \tag{16}$$

The Continuous Ranked Probability Score

If the forecast cdf $F(x)$ is Normal with mean μ and variance σ^2, then the CRPS is given by [Gneiting et al., 2005]

$$crps(F,y) = \sigma\left\{\frac{y-\mu}{\sigma}\left[2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1\right] + 2\varphi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right\}. \tag{17}$$

Grimit et al. [2006] showed that the CRPS of a mixture of Normal distributions is given by

$$crps(F,y) = \frac{1}{K}\sum_{k=1}^{K}\omega_k A\left(y-\mu_k, \sigma_k^2\right) - \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K}\omega_k\omega_l A\left(\mu_k - \mu_l, \sigma_k^2 + \sigma_l^2\right), \tag{18}$$

where

$$A\left(\mu,\sigma^2\right) = 2\sigma\varphi\left(\frac{\mu}{\sigma}\right) + \mu\left(2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right) \tag{19}$$

Analytical results for the CRPS of other distributions are difficult to derive. For example, no result has been published for the CRPS of the t-distribution, which appears as a predictive distribution in Equation 6. Numerical integration can be used where analytic results do not exist. The CRPS of forecasts issued as t-distributions was calculated using the function `integrate` as implemented in package `stats` of the `R` statistical computing software [R Core Team, 2015, version 3.2.0].

Probability Integral Transformations

If the forecast pdf $f_t(x)$ is Normal with mean μ_t and standard deviation σ_t, then the PIT is given by

$$pit(f_t, y_t) = \Phi\left(\frac{y_t - \mu_t}{\sigma_t}\right). \tag{20}$$

The PIT of a weighted sum of Normals is equal to the weighted sum of the PITs of the individual Normals.

If the forecast pdf $f_t(x)$ is issued as a non-standardized t-distribution with location μ_t, scale σ_t, and ν_t degrees-of-freedom, then the PIT is given by

$$pit\left(f_t, y_t\right) = \mathrm{T}_{\nu_t}\left(\frac{y_t - \mu_t}{\sigma_t}\right), \tag{21}$$

where $T_\nu(x)$ is the cdf of the central t-distribution with ν degrees-of-freedom.

References

J. Bröcker and L. A. Smith. From ensemble forecasts to predictive distribution functions. *Tellus A*, 60(4):663–678, Aug 2008. doi: 10.1111/j.1600-0870.2008.00333.x.

P. Brohan, J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research: Atmospheres*, 111(D12), 2006. doi: 10.1029/2005JD006548. D12106.

A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

D. Dee, S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, P. Bauer, et al. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597, 2011. doi: 10.1002/qj.828.

N. R. Draper, H. Smith, and E. Pownell. *Applied regression analysis*, volume 3. Wiley New York, 1998.

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, August 2014. doi: 10.1002/2014GL061146.

B. Efron. *The Jackknife, the Bootstrap, and Other Resampling Plans*, volume 38. SIAM, 1982.

B. Glahn, M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, and B. Schuknecht. MOS uncertainty estimates in an ensemble framework. *Monthly Weather Review*, 137(1):246–268, Jan 2009a. doi: 10.1175/2008MWR2569.1.

B. Glahn, M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht, and B. Jackson. MOS uncertainty estimates in an ensemble framework. *Monthly Weather Review*, 137(1):246–268, Jan 2009b. doi: 10.1175/2008mwr2569.1.

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. doi: 10.1175/1520-0450(1972)011⟨1203:tuomos⟩2.0.co;2.

T. Gneiting, A. E. Raftery, A. H. Westveld, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, May 2005. doi: 10.1175/mwr2904.1.

T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

E. P. Grimit, T. Gneiting, V. J. Berrocal, and N. A. Johnson. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Q.J.R. Meteorol. Soc.*, 132(621C):2925–2942, Oct 2006. doi: 10.1256/qj.05.235.

R. Hagedorn, T. M. Hamill, and J. S. Whitaker. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. part I: Two-meter temperatures. *Monthly Weather Review*, 136(7):2608–2619, Jul 2008. doi: 10.1175/2007mwr2410.1.

T. M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560, Mar 2001. doi: 10.1175/1520-0493(2001)129⟨0550:iorhfv⟩2.0.co;2.

T. M. Hamill, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu, and W. Lapenta. NOAAs second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society*, 94(10):1553–1565, Oct 2013. doi: 10.1175/bams-d-12-00014.1.

I. R. Harris. Predictive fit for natural exponential families. *Biometrika*, 76(4):675–684, 1989. doi: 10.1093/biomet/76.4.675.

H. Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.

A. Kann, C. Wittmann, Y. Wang, and X. Ma. Calibrating 2-m temperature of limited-area ensemble forecasts using high-resolution analysis. *Monthly Weather Review*, 137(10):3373–3387, Oct 2009. doi: 10.1175/2009mwr2793.1.

V. V. Kharin and F. W. Zwiers. Improved seasonal probability forecasts. *J. Climate*, 16(11):16840–1701, Jun 2003. doi: 10.1175/1520-0442(2003)016⟨1684:ispf⟩2.0.co;2.

C. E. Leith. Theoretical skill of monte carlo forecasts. *Monthly Weather Review*, 102(6):409–418, Jun 1974. doi: 10.1175/1520-0493(1974)102⟨0409:tsomcf⟩2.0.co;2.

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. Marty, V. Fortin, H. Kuswanto, A.-C. Favre, and E. Parent. Combining the Bayesian processor of output with Bayesian model averaging for reliable ensemble forecasting. *J. R. Stat. Soc. C*, 64(1):75–92, May 2014. doi: 10.1111/rssc.12062.

S. J. Mason and G. M. Mimmack. Comparison of some statistical methods of probabilistic forecasting of ENSO. *J. Climate*, 15(1):8–29, Jan 2002. doi: 10.1175/1520-0442(2002)015⟨0008:cossmo⟩2.0.co;2.

J. E. Matheson and R. L. Winkler. Scoring rules for continuous probability distributions. *Management science*, 22(10):1087–1096, 1976.

W. Merryfield, B. Denis, J. Fontecilla, W. Lee, V. Kharin, J. Hodgson, and B. Archambault. The Canadian seasonal to interannual prediction system (CanSIPS): an overview of its design and operational implementation. CMC Technical Report `http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/lib/op_systems/doc_opchanges` 2011. (Online, accessed 23 July 2015).

F. Molteni, T. Stockdale, M. Balmaseda, G. Balsamo, R. Buizza, L. Ferranti, L. Magnusson, K. Mogensen, T. Palmer, and F. Vitart. The new ECMWF seasonal forecast system (System 4), 2011. ECMWF Technical Memorandum No. 656 URL `http://old.ecmwf.int/publications/library/ecpublications/_pdf/tm/601-700/tm656.pdf`.

M. Pagowski, G. Grell, D. Devenyi, S. Peckham, S. McKeen, W. Gong, L. Delle Monache, J. McHenry, J. McQueen, and P. Lee. Application of dynamic linear regression to improve the skill of ensemble-based deterministic ozone forecasts. *Atmospheric Environment*, 40(18):3240–3250, Jun 2006. ISSN 1352-2310.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6):1653–1660, Jun 2002. doi: 10.1175/1520-0493(2002)130⟨1653:epfuit⟩2.0.co;2.

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.

S. Siegert, D. B. Stephenson, P. G. Sansom, A. A. Scaife, R. Eade, and A. Arribas. A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability? *J. Climate*, e-View, 2015. doi: 10.1175/JCLI-D-15-0196.1.

K.-T. Sohn, D.-K. Rha, and Y.-K. Seo. The 3-hour-interval prediction of ground-level temperature in South Korea using dynamic linear models. *Advances in Atmospheric Sciences*, 20(4):575–582, Jul 2003. doi: 10.1007/bf02915500.

M. K. Tippett, L. Goddard, and A. G. Barnston. Statistical-dynamical seasonal forecasts of central-southwest Asian winter precipitation. *Journal of climate*, 18(11):1831–1843, 2005.

D. A. Unger, H. van den Dool, E. OLenic, and D. Collins. Ensemble regression. *Monthly Weather Review*, 137(7):2365–2379, Jul 2009. doi: 10.1175/2008mwr2605.1.

M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, 2nd edition, 1997.

R. Williams, C. Ferro, and F. Kwasniok. A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140(680):1112–1120, 2014.