Bayesian Ensemble Model Output Statistics for Temperature

Diplomarbeit
von
Dina Richter

Betreuer:
Prof. Dr. Tilmann Gneiting
Dr. Alex Lenkoski
Dr. Thordis L. Thorarinsdottir

Ruprecht-Karls-Universität Heidelberg
Fakultät für Mathematik und Informatik
Juni 2012

Zusammenfassung

In der vorliegenden Arbeit schlagen wir ein Bayessches Modell vor, das probabilistische Vorhersagen für die stetige Größe Temperatur hervorbringt. Für die Vorhersagen werden Normalverteilungen benutzt, mit einem Erwartungswert, der linear ist in einer unbekannten Parametermenge, und mit einer unbekannten Varianz. Dabei ist der Erwartungswert ein gewichteter Durchschnitt von deterministischen Vorhersagen eines Ensembles. Die Ungewissheit und Unkenntnis der Parameter wird abgebildet, indem Wahrscheinlichkeitsverteilungen gebildet werden. Erste Vorstellungen über diese Werte werden in die Priori-Verteilungen des Erwartungswertes und der Varianz einbezogen. Nach Berücksichtigung von Trainingsdaten werden unsere Annahmen über Parameterschätzungen aufdatiert, was zu korrigierten Posteriori-Verteilungen führt. Mit deren Hilfe erstellen wir ein Vorhersagemodell und erzeugen Vorhersagen für Testdaten. Bei der Anwendung unserer Bayesschen Methode auf das University of Washington Mesoscale Ensemble über dem nordamerikanischen Pazifischen Nordwesten erhalten wir kalibrierte Vorhersagen und gute Ergebnisse.

Abstract

In this study, we propose a Bayesian model which produces probabilistic forecasts for the continuous weather variable temperature. The predictive distributions are Gaussian with a mean that is linear in an unknown parameter set, and an unknown variance. Furthermore, the mean is a bias-corrected weighted average of deterministic ensemble forecasts. The uncertainty and ignorance of the parameters is expressed by forming probability distributions. Prior beliefs are included in our prior distributions for the mean and variance. After considering training data, our assumptions of the parameter estimates are improved leading to posterior distributions. With their help, we build a predictive model to create forecasts for the test data. When applying our Bayesian method to the University of Washington mesoscale ensemble over the North American Pacific Northwest, we obtain calibrated forecasts and good performance under proper scoring rules.

Contents

1 Introduction

In Numerical Weather Prediction (NWP) methods, forecast models produce deterministic predictions for weather quantities. The model system depends on the initial or boundary conditions which are given by an estimate of the current state of the atmosphere, and even if the measurements do not vary significantly, the resulting predictions may be very different due to the different representation of the atmospheric processes. Furthermore, the equations of the atmospheric model are nontrivial and the solutions of the physical processes are approximated, leading to deficient forecasts (Britannica, 2012). The approximations in the numerics and the input data tend to result in predictions that are not completely accurate. An ensemble consists of multiple NWPs for a single variable. The future atmospheric states are here generated with differing initial conditions, and may result from one or several models. Forecasting using ensembles usually leads to improvement, since the prediction uncertainty is included.

Statistical postprocessing techniques then link the results of the numerical forecasts to statistical models resulting in improved predictions. One such method is the Model Output Statistics (MOS) technique of Glahn and Lowry (1972), where regression equations produce forecasts of surface weather variables.

We concentrate on daily prediction of the continuous weather quantity temperature, and our method, which is an extension of the Ensemble Model Output Statistics (EMOS) method of Gneiting et al. (2005), yields probabilistic forecasts in form of full predictive distributions within a fully Bayesian framework.

The remainder of this thesis is organized as follows. In Chapter 2, we summarize other postprocessing methods that have also dealt with temperature forecasts and outline our Bayesian approach. A detailed description of our method follows in the next Chapter, where we also state how the skill of the models can be measured. In Chapter 4, our results are reported, where we have applied our Bayesian approach and competing forecasting procedures to make temperature forecasts over the North-American Pacific Northwest in the year 2008 using the University of Washington Mesoscale Ensemble (UWME). Finally, conclusions are provided in Chapter 5.

2 Probabilistic forecasts for temperature

2.1 EMOS

The postprocessing technique Ensemble Model Output Statistics (EMOS) (Gneiting et al., 2005) uses an ensemble consisting of distinguishable forecasts for some univariate weather quantity such as surface temperature or sea pressure. The method is used to correct for forecast bias and underdispersion of the ensemble, and it is based on multiple linear regression. Its probabilistic forecasts have Gaussian predictive distributions, which are the EMOS forecasts.

The mean of the predictive distribution is a weighted average of the ensemble member forecasts, corrected for bias. That is, let the ensemble be $\mathbf{X} = \{X_1, \ldots, X_m\}$, then the mean is modeled as

$$\mu = a + b_1 X_1 + \ldots + b_m X_m,$$

where a is a bias-correction, and $b_1, \ldots, b_m$ are regression coefficients, which show the skill of the members over a training set. The variance depends linearly on the ensemble variance and takes the spread-skill relationship into account. If c and d are nonnegative numbers, and S^2 denotes the ensemble variance, the variance has the form

$$\sigma^2 = c + dS^2.$$

For the weather variable Y, it follows that

$$Y|\mathbf{X} \sim \mathcal{N}\left(\mu, \sigma^2\right).$$

To find the EMOS coefficients, the authors introduce the method of minimum Continuous Ranked Probability Score (CRPS) estimation which is applied to the training data. Furthermore, a variant of the technique is used, EMOS$^+$, where the coefficients are constrained to be nonnegative. In the first step of EMOS$^+$, the coefficients of the EMOS model are estimated by optimizing the CRPS. If all coefficients are nonnegative, then both the EMOS and the EMOS$^+$ forecasts are the same. However, if one or several

values are negative, then these regression coefficients are set to zero. The CRPS optimum now has to be found using the reduced ensemble. This procedure is repeated until all estimated parameters are nonnegative. In each step, the ensemble variance has to be calculated anew using the remaining ensemble members.

The authors state that the removed ensemble members of the model EMOS$^+$ are less useful relative to the included members over the training period. However, the EMOS and EMOS$^+$ forecasts were found to be equally skillful.

2.2 Ensemble BMA

Raftery et al. (2005) propose an alternative ensemble postprocessing method. Their method, Bayesian Model Averaging (BMA), combines predictive distributions from different competing models. In this approach, the predictive probability distribution function (PDF) of the weather quantity (temperature or sea level pressure) is a weighted average of PDFs based on individual forecasts.

The weights are the estimated posterior probabilities of the competing ensemble members and show the predictive performance of each member relative to the other members over the training set. Small values of the weights indicate that the corresponding ensemble member was less useful for the training period.

A future observation is denoted by y, and y_T is the training data. If K different models $M_1, \ldots, M_K$ are considered, then the predictive PDF for y is

$$p(y) = \sum_{k=1}^{K} p(y|M_k)\, p(M_k|y_T),$$

where $p(y|M_k)$ is the PDF conditionally on model M_k, and $p(M_k|y_T)$ denotes the posterior probability of model M_k given the training data set y^T. Model M_k conditions y on ensemble member f_k and its PDF is a normal distribution with mean $a_k + b_k f_k$, and variance σ^2:

$$y|f_k \sim \mathcal{N}\left(a_k + b_k f_k, \sigma^2\right).$$

Then the BMA model is

$$p(y|f_1, \ldots, f_k) = \sum_{k=1}^{K} w_k \mathcal{N}\left(a_k + b_k f_k, \sigma^2\right),$$

where w_k is the posterior probability that f_k is the best forecast depending on how it performed in the training set, and which is estimated using maximum likelihood. This

method produces calibrated and sharp predictive PDF.

2.3 A Bayesian approach to ensemble BMA

A Bayesian approach to the statistical postprocessing of temperature forecasts was proposed by Narzo and Cocchi (2010). Here, the ensemble members are exchangeable. That is, the system produces ensemble members that are viewed as random replications of the same data-generating process.

Let the observations be denoted by y_t, and the K forecast ensemble members by $X_t = \{X_{tk}, k = 1, \ldots, K\}$. A latent process selects a member from the ensemble and a Bayesian hierarchical model is used to relate an observation with just this ensemble member and not the full ensemble.

In the first level, a distribution for observed values is conditioned on the selected deterministic forecast. That is, a chosen ensemble member x_t on day t can be related to y_t: The distribution is Gaussian, with a mean that is a linear function of the forecast x_{ts}, where the index s denotes a particular station. This results in

$$y_{ts}|x_{ts} \sim \mathcal{N}\left(\alpha_s + \beta x_{ts}, \sigma_y^2\right),$$

where β is a common slope for all stations, α_s is a station-specific intercept, and σ_y^2 is a common error variance. The intercept α_s is normally distributed, so that

$$\alpha_s \sim \mathcal{N}\left(\alpha_0, \sigma_\alpha^2\right).$$

In the second level, the selection process with the outcome of one of the K ensemble members is modeled on each day t.

Given training data, a predictive probability distribution of new data can produce forecasts.

2.4 This study: BEMOS

To adopt a Bayesian statistical method, beliefs about unknown parameters are formed by assigning probabilities to them, expressing the uncertainty about the true parameter values. The application of Bayes' Theorem leads to an update of those beliefs using the given information. This growing knowledge about the parameters is called Bayesian inference.

Let the parameter θ contained in the parameter set Θ express the unknown properties of a quantity we are interested in. For every $\theta \in \Theta$ the prior distribution $p(\theta)$ expresses the belief that θ is the true value. For every $\theta \in \Theta$ and every dataset y of the sample space, the sampling model $p(y|\theta)$ characterizes the results of the set y under the assumption that θ is true. The information from the observed data is then used to update the uncertainty about θ, so that for every $\theta \in \Theta$ the posterior distribution $p(\theta|y)$ expresses the updated belief about θ. To obtain the posterior distribution, Bayes' Theorem yields

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{p(y)} \propto p(y|\theta)\, p(\theta).$$

In our method, the sample model for the data depends on covariates consisting of an eight-member ensemble forecast, and we want to formulate a prediction model for temperature. For that reason, training data y will be used to estimate parameters in a regression model similar to the EMOS model in Chapter 2.1 and after that, the model is assessed using a test dataset $\tilde{y}$. The predictive distribution of the new observations is conditioned on the observed data and has the following form:

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y)\, d\theta \\ &= \int p(\tilde{y}|\theta)\, p(\theta|y)\, d\theta. \end{aligned}$$

As the sample model is Gaussian, the weather quantity $\tilde{y}$ has a Gaussian predictive distribution, with a mean that is linear in the ensemble members and a variance σ^2, so that the unknown parameter set θ consists of the regression coefficients and σ^2. We assess the performance of our Bayesian ensemble model output statistics (BEMOS) for forecasting temperature 48-h ahead using the University of Washington mesoscale ensemble over the North-American Pacific Northwest in 2008 and compare the results to the EMOS method of Gneiting et al. (2005).

3 Methods

3.1 BEMOS

In our study, we largely follow Hoff (2009). We consider both the so-called g-prior and other weakly informative prior for the regression coefficients and show how samples from the posterior distributions may be obtained using Gibbs sampling. Finally, we discuss model uncertainty, where a Bayesian model selection procedure over all possible regression models is presented.

3.1.1 Linear regression

We provide a brief overview of linear regression analysis, which is the basis of our Bayesian approach. In a regression model, we deal with a set of explanatory variables or regressors $\mathbf{x} = (x_1, \ldots, x_p)$. The distribution of a random variabe Y depends on this set and we write $p(y|\mathbf{x})$, which denotes the conditional distribution of Y given $\mathbf{x}$. A linear regression model is characterized by the assumption that the expectation of Y given $\mathbf{x}$ is linear in a parameter set $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, that is,

$$\mathbb{E}[Y|\mathbf{x}] = \int y\, p(y|\mathbf{x})\, dy = \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{\beta}^T \mathbf{x}.$$

Additionally, the first variable x_1 is often set as $x_1 = 1$, so that β_1 corresponds to a bias-correction term.

In a normal linear regression model, Y varies around the mean $\mathbb{E}[Y|\mathbf{x}]$ with an independent error term that follows a normal distribution. For $i = 1, \ldots, n$ the random variable Y_i can be expressed as follows:

$$\begin{aligned} \varepsilon_i &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \\ Y_i &= \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i. \end{aligned}$$

This characterization leads to the joint probability density of n observations, arranged in the n-dimensional vector $\mathbf{y} = (y_1, \ldots, y_n)^T$. The conditions $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are summarized

as rows of an $n \times p$ matrix $\mathbf{X}$ and we write

$$\begin{aligned} p(y_1, \ldots, y_n | \mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} p(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 \right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right) \\ &= p(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2). \end{aligned}$$

The expression in the exponent, $\sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2$, is the sum of squared residuals and will be referred to as SSR($\boldsymbol{\beta}$) .

With this notation, $\mathbf{y}$ conditional on $\mathbf{X}, \boldsymbol{\beta}$ and σ^2 follows a multivariate normal distribution,

$$\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbb{I}),$$

where $\mathbb{I}$ is the $n \times n$ identity matrix and

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \beta_1 x_{1,1} + \cdots + \beta_p x_{1,p} \\ \vdots \\ \beta_1 x_{n,1} + \cdots + \beta_p x_{n,p} \end{pmatrix} = \begin{pmatrix} \mathbb{E}[Y_1 | \boldsymbol{\beta}, \mathbf{x}_1] \\ \vdots \\ \mathbb{E}[Y_n | \boldsymbol{\beta}, \mathbf{x}_n] \end{pmatrix}.$$

The parameters to be estimated are thus $\theta = (\boldsymbol{\beta}, \sigma^2)$.

3.1.2 Bayesian parameter estimation

For computational convenience, we aim to define the prior distributions for the parameters $\boldsymbol{\beta}$ and σ^2 in such a way that the joint posterior distribution $p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X})$ may be approximated using a Gibbs sampler.

If the prior density of $\boldsymbol{\beta}$ is a multivariate normal distribution,

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\beta}_0, \Sigma_0),$$

then it is a conjugate prior leading to a multivariate normal posterior distribution as well. This will be shown in the following.

The posterior of $\boldsymbol{\beta}$ is proportional to $p(\boldsymbol{\beta})\, p(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2)$, and only the terms depending

on $\boldsymbol{\beta}$ need to be considered. For the sampling density of the data, we write:

$$\begin{aligned} p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2) &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\mathrm{SSR}(\boldsymbol{\beta})\right) \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y}-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta})\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta})\right). \end{aligned}$$

Similar results hold for the prior distribution $p(\boldsymbol{\beta})$ and it follows for the full conditional posterior dstribution for $\boldsymbol{\beta}$ that

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2) \propto& p(\boldsymbol{\beta})\,p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2) \\ \propto& \exp\left(-\frac{1}{2}(-2\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}_0+\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta})-\frac{1}{2\sigma^2}(-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta})\right) \\ =& \exp\left(-\frac{1}{2}[\boldsymbol{\beta}^T(\Sigma_0^{-1}+\mathbf{X}^T\mathbf{X}/\sigma^2)\boldsymbol{\beta}-2\,\boldsymbol{\beta}^T(\Sigma_0^{-1}\boldsymbol{\beta}_0+\mathbf{X}^T\mathbf{y}/\sigma^2)]\right). \end{aligned}$$

As a consequence, we see that the posterior distribution is proportional to a multivariate normal density, i. e. $\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2 \sim \mathcal{N}_p(\mathbf{m},\mathbf{V})$, where

$$\begin{aligned} \mathbf{m} = \mathbb{E}[\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2] \quad &= (\Sigma_0^{-1}+\mathbf{X}^T\mathbf{X}/\sigma^2)^{-1}(\Sigma_0^{-1}\boldsymbol{\beta}_0+\mathbf{X}^T\mathbf{y}/\sigma^2), \\ \mathbf{V} = \mathbb{V}\mathrm{ar}[\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2] &= (\Sigma_0^{-1}+\mathbf{X}^T\mathbf{X}/\sigma^2)^{-1}. \end{aligned}$$

For the prior distribution for σ^2, we choose an inverse-gamma distribution,

$$\frac{1}{\sigma^2} \sim \Gamma\left(\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2}\right),$$

which results in the full conditional posterior

$$\begin{aligned} p(\tfrac{1}{\sigma^2}|\mathbf{y},\mathbf{X},\boldsymbol{\beta}) \propto& \; p(\tfrac{1}{\sigma^2})\,p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2) \\ \propto& \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0}{2}-1}\exp\left(-\frac{1}{\sigma^2}\frac{\nu_0\sigma_0^2}{2}\right)\times\left(\frac{1}{\sigma^2}\right)^{n/2}\exp\left(-\frac{1}{\sigma^2}\frac{\mathrm{SSR}(\boldsymbol{\beta})}{2}\right) \\ =& \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0+n}{2}-1}\exp\left(-\frac{1}{\sigma^2}\frac{\nu_0\sigma_0^2+\mathrm{SSR}(\boldsymbol{\beta})}{2}\right). \end{aligned}$$

This distribution is also an inverse-gamma density, that is,

$$\frac{1}{\sigma^2}|\mathbf{y}, \mathbf{X}, \boldsymbol{\beta} \sim \Gamma\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + \text{SSR}(\boldsymbol{\beta})}{2}\right).$$

With the knowledge of the full conditional posterior distributions for the parameters $\boldsymbol{\beta}$ and σ^2, the joint posterior distribution $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})$ may be approximated using a Gibbs sampler. In this process, each parameter is updated individually according to its full conditional posterior distribution, where the samples may potentially be dependent. Given the latest values $(\boldsymbol{\beta}^{(s)}, \sigma^{2(s)})$, the parameters are updated in the following way:

1. Update of $\boldsymbol{\beta}$:
   (i) compute $\mathbf{V} = \mathbb{V}\text{ar}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^{2(s)}]$ and $\mathbf{m} = \mathbb{E}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^{2(s)}]$
   (ii) sample $\boldsymbol{\beta}^{(s+1)} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{V})$
2. Update of σ^2:
   (i) compute $\text{SSR}(\boldsymbol{\beta}^{(s+1)})$
   (ii) sample $\left(\frac{1}{\sigma^2}\right)^{(s+1)} \sim \Gamma\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2+\text{SSR}(\boldsymbol{\beta}^{(s+1)})}{2}\right)$.

3.1.3 g-prior distribution

A weakly informative prior distribution for regression coefficients is the so-called g-prior (Zellner, 1986), with

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\beta}_0, g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}),$$

where $g > 0$. Such a construction simplifies the terms $\mathbf{m}$ and $\mathbf{V}$ of the multivariate normal posterior distribution for $\boldsymbol{\beta}$, where now it holds that

$$\begin{aligned}
\mathbb{V}\text{ar}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] &= (\tfrac{1}{g\sigma^2}\mathbf{X}^T\mathbf{X} + \tfrac{1}{\sigma^2}\mathbf{X}^T\mathbf{X})^{-1} \\
&= \frac{g}{g+1}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}, \\
\mathbb{E}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] &= (\tfrac{1}{g\sigma^2}\mathbf{X}^T\mathbf{X} + \tfrac{1}{\sigma^2}\mathbf{X}^T\mathbf{X})^{-1}(\tfrac{1}{g\sigma^2}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0 + \tfrac{1}{\sigma^2}\mathbf{X}^T\mathbf{y}) \\
&= \frac{1}{g+1}\boldsymbol{\beta}_0 + \frac{g}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.
\end{aligned}$$

Under the g-prior, the marginal posterior density of σ^2 conditional on $\mathbf{y}$ and $\mathbf{X}$ may be calculated explicitly. As a prior for σ^2, we still take the inverse-gamma distribution,

$$\frac{1}{\sigma^2} \sim \Gamma\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right).$$

The marginal posterior density is proportional to $p(\sigma^2)\,p(\mathbf{y}|\mathbf{X},\sigma^2)$, where

$$p(\mathbf{y}|\mathbf{X},\sigma^2) = \int p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2)\,p(\boldsymbol{\beta}|\mathbf{X},\sigma^2)\,d\boldsymbol{\beta}.$$

The integrand is

$$\begin{aligned} p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2)\,p(\boldsymbol{\beta}|\mathbf{X},\sigma^2) =& (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)\times \\ & |2\pi g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}|^{-1/2}\exp\left(-\frac{1}{2g\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right), \end{aligned}$$

where the exponents can be rewritten as follows:

$$\begin{aligned} & -\frac{1}{2\sigma^2}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)/g\right] \\ =& -\frac{1}{2\sigma^2}\left[\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + (\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\beta}_0^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0)/g\right]. \end{aligned}$$

Ignoring the terms in the brackets depending on $\mathbf{y}$, and rearranging the remaining expressions, we obtain

$$\begin{aligned} & \frac{g+1}{g}\left[\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0/(g+1) + \boldsymbol{\beta}_0^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0/(g+1)^2\right] \\ & -\frac{g+1}{g}\boldsymbol{\beta}_0^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0/(g+1)^2 + \frac{1}{g}\boldsymbol{\beta}_0^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0 \\ =& \frac{g+1}{g}(\boldsymbol{\beta} - \tfrac{1}{g+1}\boldsymbol{\beta}_0)^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta} - \tfrac{1}{g+1}\boldsymbol{\beta}_0) + \frac{1}{g+1}\boldsymbol{\beta}_0^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0. \end{aligned}$$

It follows for the exponent that

$$\begin{aligned} & -\frac{1}{2\sigma^2}\left[\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \frac{g+1}{g}(\boldsymbol{\beta} - \tfrac{1}{g+1}\boldsymbol{\beta}_0)^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta} - \tfrac{1}{g+1}\boldsymbol{\beta}_0) + \frac{1}{g+1}\boldsymbol{\beta}_0^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0\right] \\ =& -\frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}(\boldsymbol{\beta}-\widetilde{\mathbf{m}})^T\mathbf{V}^{-1}(\boldsymbol{\beta}-\widetilde{\mathbf{m}}) + \frac{1}{2}\mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} + \frac{1}{\sigma^2(g+1)}\mathbf{y}^T\mathbf{X}\boldsymbol{\beta}_0 \\ & -\frac{1}{2\sigma^2(g+1)}\boldsymbol{\beta}_0^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0, \end{aligned}$$

where

$$\widetilde{\mathbf{m}} = \frac{1}{g+1}\boldsymbol{\beta}_0 + \frac{g}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \frac{1}{g+1}\boldsymbol{\beta}_0 + \mathbf{m} \quad \text{and} \quad \mathbf{V} = \frac{g}{g+1}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Inserting this term into the integrand of $p(\mathbf{y}|\mathbf{X}, \sigma^2)$ results in

$$
\begin{aligned}
& p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta}|\mathbf{X}, \sigma^2) \\
= & \left[|2\pi\mathbf{V}|^{-1/2} \exp[-\frac{1}{2}(\boldsymbol{\beta} - \widetilde{\mathbf{m}})^T \mathbf{V}^{-1} (\boldsymbol{\beta} - \widetilde{\mathbf{m}})] \right] \\
& \times \left[(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \mathbf{y}^T \mathbf{y} \right) \right] \\
& \times \left[(1+g)^{-p/2} \exp\left(\frac{1}{2} \mathbf{m}^T \mathbf{V}^{-1} \mathbf{m} + \frac{1}{\sigma^2(g+1)} \mathbf{y}^T \mathbf{X} \boldsymbol{\beta}_0 - \frac{1}{2\sigma^2(g+1)} \boldsymbol{\beta}_0^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}_0 \right) \right].
\end{aligned}
$$

This expression has to be integrated with respect to $\boldsymbol{\beta}$, where only the first term depends on $\boldsymbol{\beta}$. The term that depends on $\boldsymbol{\beta}$ is the multivariate normal density with mean $\widetilde{\mathbf{m}}$ and variance $\mathbf{V}$, and it thus integrates to 1. Therefore, only the latter two terms remain. As a consequence,

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{X}, \sigma^2) &= \int p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta}|\mathbf{X}, \sigma^2)\, \mathrm{d}\boldsymbol{\beta} \\
&= \left[(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \mathbf{y}^T \mathbf{y} \right) \right] \times (1+g)^{-p/2} \\
&\quad \times \exp\left(\frac{1}{2} \mathbf{m}^T \mathbf{V}^{-1} \mathbf{m} + \frac{1}{\sigma^2(g+1)} \mathbf{y}^T \mathbf{X} \boldsymbol{\beta}_0 - \frac{1}{2\sigma^2(g+1)} \boldsymbol{\beta}_0^T \mathbf{X}_0^T \mathbf{X} \boldsymbol{\beta}_0 \right) \\
&= (2\pi)^{-n/2} (1+g)^{-p/2} (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \mathrm{SSR}_g \right),
\end{aligned}
$$

where SSR_g is

$$
\begin{aligned}
\mathrm{SSR}_g &= \mathbf{y}^T \mathbf{y} - \sigma^2 \mathbf{m}^T \mathbf{V}^{-1} \mathbf{m} - \frac{2}{g+1} \mathbf{y}^T \mathbf{X} \boldsymbol{\beta}_0 + \frac{1}{g+1} \boldsymbol{\beta}_0^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}_0 \\
&= \mathbf{y}^T (\mathbb{I} - \frac{g}{g+1} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{y} - \frac{2}{g+1} \mathbf{y}^T \mathbf{X} \boldsymbol{\beta}_0 + \frac{1}{g+1} \boldsymbol{\beta}_0^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}_0.
\end{aligned}
$$

Therefore, the posterior distribution of σ^2 is given by

$$
\begin{aligned}
p(\tfrac{1}{\sigma^2}|\mathbf{y}, \mathbf{X}) &\propto p(\tfrac{1}{\sigma^2})\, p(\mathbf{y}|\mathbf{X}, \sigma^2) \\
&\propto \left(\frac{1}{\sigma^2} \right)^{\frac{\nu_0}{2}-1} \exp\left(-\frac{1}{\sigma^2} \frac{\nu_0 \sigma_0^2}{2} \right) \times \left(\frac{1}{\sigma^2} \right)^{n/2} \exp\left(-\frac{1}{\sigma^2} \frac{\mathrm{SSR}_g}{2} \right) \\
&= \left(\frac{1}{\sigma^2} \right)^{\frac{\nu_0+n}{2}-1} \exp\left(-\frac{1}{\sigma^2} \frac{\nu_0 \sigma_0^2 + \mathrm{SSR}_g}{2} \right),
\end{aligned}
$$

which is also proportional to an inverse-gamma density with

$$\frac{1}{\sigma^2}|\mathbf{y}, \mathbf{X} \sim \Gamma\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + \text{SSR}_g}{2}\right).$$

As the posterior of σ^2 does not depend on $\boldsymbol{\beta}$, we may sample from the joint posterior distribution $p(\sigma^2, \boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ with a Monte Carlo approximation. A sample $(\sigma^2, \boldsymbol{\beta})$ is obtained by

1. sampling $\frac{1}{\sigma^2}|\mathbf{y}, \mathbf{X} \sim \Gamma\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2+\text{SSR}_g}{2}\right)$
2. then sampling $\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2 \sim \mathcal{N}\left(\frac{1}{g+1}\boldsymbol{\beta}_0 + \frac{g}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \frac{g}{g+1}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$.

This is a generalization of the model discussed in Hoff (2009) who only considers $\boldsymbol{\beta}_0 = \mathbf{0}$. In our case study, each ensemble member is a priori equally likely, so that $\boldsymbol{\beta}_0 = (0, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8)$.

3.1.4 Weakly informative prior

As a further option, we propose a weakly informative prior for $\boldsymbol{\beta}$ of the form

$$\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}_p\left(\boldsymbol{\beta}_0, \frac{\sigma^2}{n_0}\mathbb{I}_p\right).$$

Similar to the g-prior, the factors of the multivariate normal posterior distribution $\mathbf{m}$ and $\mathbf{V}$ are simplified compared to the multivariate prior in Section 3.1.2. Here, we obtain

$$\begin{aligned}
\mathbb{V}\text{ar}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] &= (\tfrac{n_0}{\sigma^2}\mathbb{I}_p + \tfrac{1}{\sigma^2}\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(n_0\mathbb{I}_p + \mathbf{X}^T\mathbf{X})^{-1} =: \sigma^2\tilde{\Sigma}, \\
\mathbb{E}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] &= (\tfrac{n_0}{\sigma^2}\mathbb{I}_p + \tfrac{1}{\sigma^2}\mathbf{X}^T\mathbf{X})^{-1}(\tfrac{n_0}{\sigma^2}\mathbb{I}_p\boldsymbol{\beta}_0 + \tfrac{1}{\sigma^2}\mathbf{X}^T\mathbf{y}) \\
&= \tilde{\Sigma}(n_0\boldsymbol{\beta}_0 + \mathbf{X}^T\mathbf{y}) =: \tilde{\boldsymbol{\beta}}.
\end{aligned}$$

The prior distribution for σ^2 continues to be the inverse-gamma distribution,

$$\frac{1}{\sigma^2} \sim \Gamma\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right).$$

For σ^2, the marginal posterior density is proportional to $p(\sigma^2)\, p(\mathbf{y}|\mathbf{X}, \sigma^2)$, where here

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \sigma^2) &= \int p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta}|\mathbf{X}, \sigma^2)\, d\boldsymbol{\beta} \\ &= \int (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \times \\ &\qquad |2\pi\tfrac{\sigma^2}{n_0}\mathbb{I}_p|^{-1/2} \exp\left(-\frac{n_0}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right)\, d\boldsymbol{\beta}. \end{aligned}$$

Rearranging the exponents in the integrand results in

$$\begin{aligned} & -\frac{1}{2\sigma^2}\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + n_0(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right] \\ =& -\frac{1}{2\sigma^2}\left[\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + n_0\boldsymbol{\beta}^T\boldsymbol{\beta} - 2n_0\boldsymbol{\beta}^T\boldsymbol{\beta}_0 + n_0\boldsymbol{\beta}_0^T\boldsymbol{\beta}_0\right] \\ =& -\frac{1}{2\sigma^2}\left[\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T(n_0\mathbb{I}_p + \mathbf{X}^T\mathbf{X})\boldsymbol{\beta} - 2n_0\boldsymbol{\beta}^T\boldsymbol{\beta}_0 + n_0\boldsymbol{\beta}_0^T\boldsymbol{\beta}_0\right] \\ =& -\frac{1}{2\sigma^2}\left[\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\tilde{\Sigma}^{-1}\boldsymbol{\beta} - 2n_0\boldsymbol{\beta}^T\boldsymbol{\beta}_0 + n_0\boldsymbol{\beta}_0^T\boldsymbol{\beta}_0\right] \\ =& -\frac{1}{2\sigma^2}\left[\mathbf{y}^T\mathbf{y} + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T\tilde{\Sigma}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) - \tilde{\boldsymbol{\beta}}^T\tilde{\Sigma}^{-1}\tilde{\boldsymbol{\beta}} + n_0\boldsymbol{\beta}_0^T\boldsymbol{\beta}_0\right], \end{aligned}$$

where

$$\tilde{\Sigma} = (n_0\mathbb{I}_p + \mathbf{X}^T\mathbf{X})^{-1} \quad \text{and} \quad \tilde{\boldsymbol{\beta}} = \tilde{\Sigma}(n_0\boldsymbol{\beta}_0 + \mathbf{X}^T\mathbf{y}).$$

Inserting this term into the integrand of $p(\mathbf{y}|\mathbf{X}, \sigma^2)$, we get

$$\begin{aligned} & p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta}|\mathbf{X}, \sigma^2) \\ =& \left[|2\pi\sigma^2\tilde{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T\tilde{\Sigma}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\right)\right] \\ & \times \left[(2\pi\sigma^2)^{-n/2}|\sigma^2\tilde{\Sigma}|^{1/2}|\tfrac{\sigma^2}{n_0}\mathbb{I}_p|^{-1/2} \exp\left(-\frac{1}{2\sigma^2}[\mathbf{y}^T\mathbf{y} - \tilde{\boldsymbol{\beta}}^T\tilde{\Sigma}^{-1}\tilde{\boldsymbol{\beta}} + n_0\boldsymbol{\beta}_0^T\boldsymbol{\beta}_0]\right)\right]. \end{aligned}$$

Only the first term of this expression depends on $\boldsymbol{\beta}$ and as it is the multivariate normal density with mean $\widetilde{\boldsymbol{\beta}}$ and variance $\sigma^2\tilde{\Sigma}$, merely the second term is left,

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \sigma^2) &= \int p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}|\mathbf{X}, \sigma^2)\, \mathrm{d}\boldsymbol{\beta} \\ &= (2\pi\sigma^2)^{-n/2} n_0^{p/2} |\tilde{\Sigma}|^{1/2} \exp\left(-\frac{1}{2\sigma^2}[\mathbf{y}^T\mathbf{y} - \tilde{\boldsymbol{\beta}}^T\tilde{\Sigma}^{-1}\tilde{\boldsymbol{\beta}} + n_0\boldsymbol{\beta}_0^T\boldsymbol{\beta}_0]\right) \\ &= (2\pi)^{-n/2} n_0^{p/2} |\tilde{\Sigma}|^{1/2} (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\mathrm{SSR}_{n_0}\right), \end{aligned}$$

where SSR_{n_0} is given by

$$\mathrm{SSR}_{n_0} = \mathbf{y}^T\mathbf{y} - \tilde{\boldsymbol{\beta}}^T\tilde{\Sigma}^{-1}\tilde{\boldsymbol{\beta}} + n_0\boldsymbol{\beta}_0^T\boldsymbol{\beta}_0.$$

Then, the marginal posterior distribution for σ^2 is

$$\begin{aligned} p(\tfrac{1}{\sigma^2}|\mathbf{y},\mathbf{X}) &\propto p(\tfrac{1}{\sigma^2})\, p(\mathbf{y}|\mathbf{X},\sigma^2) \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0}{2}-1} \exp\left(-\frac{1}{\sigma^2}\frac{\nu_0\sigma_0^2}{2}\right) \times \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{\sigma^2}\frac{\mathrm{SSR}_{n_0}}{2}\right) \\ &= \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0+n}{2}-1} \exp\left(-\frac{1}{\sigma^2}\frac{\nu_0\sigma_0^2 + \mathrm{SSR}_{n_0}}{2}\right), \end{aligned}$$

which is proportional to an inverse-gamma density, and it holds that

$$\frac{1}{\sigma^2}|\mathbf{y},\mathbf{X} \sim \Gamma\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2 + \mathrm{SSR}_{n_0}}{2}\right).$$

As in the previous section, the posterior of σ^2 does not depend on $\boldsymbol{\beta}$. Hence, we obtain samples $(\sigma^2, \boldsymbol{\beta})$ from the joint posterior distribution $p(\sigma^2, \boldsymbol{\beta}|\mathbf{y},\mathbf{X})$ with a Monte Carlo approximation by

1. sampling $\frac{1}{\sigma^2}|\mathbf{y},\mathbf{X} \sim \Gamma\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2+\mathrm{SSR}_{n_0}}{2}\right)$
2. then sample $\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2 \sim \mathcal{N}\left(\tilde{\boldsymbol{\beta}}, \sigma^2\tilde{\Sigma}\right)$.

3.2 Incorporating model uncertainty

In the models introduced above, our random variable Y depends on regressors summarized in $\mathbf{x}$. However, it may occur that not every explanatory variable relates to the variable y. Hence, we want to consider all 2^{p-1} possible regression models for y based on different sets of regressor variables (all models include a bias-correction term). In the Bayesian procedure, the prior distribution includes the information that each regression coefficient other than the bias-correction term equals zero with positive probability. For $j = 2, \ldots, p$, the coefficient is given by

$$\beta_j = m_j b_j, \quad m_j \in \{0,1\}, b_j \in \mathbb{R},$$

while for $j = 1$, the coefficient is $\beta_1 = m_1 b_1 = b_1$ since $m_1 = 1$. The regression equation for $i = 1, \ldots, n$ becomes

$$\begin{aligned} y_i =& \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \\ =& b_1 + m_2 b_2 x_{i2} + \cdots + m_p b_p x_{ip} + \varepsilon_i. \end{aligned}$$

Then $\mathcal{M} = (m_1, \ldots, m_p)$ specifies each of the possible regression models and indicates which of the variables are included in the model by the non-zero entries. Given a prior distribution $p(\mathcal{M})$, the posterior probability for each model is

$$\begin{aligned} p(\mathcal{M}|\mathbf{y}, \mathbf{X}) &= \frac{p(\mathcal{M})\, p(\mathbf{y}|\mathbf{X}, \mathcal{M})}{\sum_{\tilde{\mathcal{M}}} p(\tilde{\mathcal{M}})\, p(\mathbf{y}|\mathbf{X}, \tilde{\mathcal{M}})} \\ &\propto p(\mathcal{M})\, p(\mathbf{y}|\mathbf{X}, \mathcal{M}), \end{aligned}$$

and if each model $\mathcal{M}$ is given an equal prior weight, we obtain

$$p(\mathcal{M}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathcal{M})}{\sum_{\tilde{\mathcal{M}}} p(\mathbf{y}|\mathbf{X}, \tilde{\mathcal{M}})}.$$

We need to determine the marginal probability $p(\mathbf{y}|\mathbf{X}, \mathcal{M})$, which can be evaluated by integrating the joint distribution of $\mathbf{y}, \boldsymbol{\beta}$ and σ^2 conditional on each model $\mathcal{M}$ with respect to the parameters $\boldsymbol{\beta}$ and σ^2,

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \mathcal{M}) &= \iint p(\mathbf{y}, \boldsymbol{\beta}, \sigma^2|\mathbf{X}, \mathcal{M})\, d\boldsymbol{\beta}\, d\sigma^2 \\ &= \iint p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}, \mathcal{M})\, p(\boldsymbol{\beta}|\sigma^2, \mathbf{X}, \mathcal{M})\, p(\sigma^2)\, d\boldsymbol{\beta}\, d\sigma^2 \\ &= \int \left(\int p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}, \mathcal{M})\, p(\boldsymbol{\beta}|\sigma^2, \mathbf{X}, \mathcal{M})\, d\boldsymbol{\beta} \right) p(\sigma^2)\, d\sigma^2 \\ &= \int p(\mathbf{y}|\sigma^2, \mathbf{X}, \mathcal{M})\, p(\sigma^2)\, d\sigma^2. \end{aligned}$$

For simplicity, we use the weakly informative prior for $\boldsymbol{\beta}$ discussed in the previous section (whereas Hoff (2009) used the g-prior distribution). For each model $\mathcal{M}$ and $j = 1, \ldots, p$, let $p_{\mathcal{M}}$ sum up the variables, where $m_j = 1$, so that $p_{\mathcal{M}} = \sum_{j=1}^{p} m_j$. Then $\mathbf{X}_{\mathcal{M}}$ defines the respective $n \times p_{\mathcal{M}}$ matrix and accordingly, $\boldsymbol{\beta}_{\mathcal{M}}$ is a vector of length $p_{\mathcal{M}}$.

Therefore, the corresponding prior distribution for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}_{\mathcal{M}}|\mathbf{X}_{\mathcal{M}}, \sigma^2 \sim \mathcal{N}_{p_{\mathcal{M}}}(\boldsymbol{\beta}_0^{\mathcal{M}}, \tfrac{\sigma^2}{n_0}\mathbb{I}_{p_{\mathcal{M}}}).$$

The marginal probability $p(\mathbf{y}|\sigma^2, \mathbf{X}, \mathcal{M})$ is known as it can be computed in the same way as shown in the previous two sections. With the following prior distribution for σ^2,

$$\frac{1}{\sigma^2} \sim \Gamma\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right),$$

it follows that

$$\begin{aligned} p(\mathbf{y}|\sigma^2, \mathbf{X}, \mathcal{M})\, p(\tfrac{1}{\sigma^2}) =& (2\pi)^{-n/2} \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{\sigma^2}\frac{\mathrm{SSR}_{n_0}^{\mathcal{M}}}{2}\right) \times n_0^{p_{\mathcal{M}}/2} |\tilde{\Sigma}_{\mathcal{M}}|^{1/2} \\ & \times \left(\frac{\nu_0\sigma_0^2}{2}\right)^{\nu_0/2} \Gamma(\tfrac{\nu_0}{2})^{-1} \left(\frac{1}{\sigma^2}\right)^{\nu_0/2-1} \exp\left(-\frac{1}{\sigma^2}\frac{\nu_0\sigma_0^2}{2}\right), \end{aligned}$$

where

$$\begin{aligned} \mathrm{SSR}_{n_0}^{\mathcal{M}} &= \mathbf{y}^T\mathbf{y} - (\tilde{\boldsymbol{\beta}}^{\mathcal{M}})^T \tilde{\Sigma}_{\mathcal{M}}^{-1} \tilde{\boldsymbol{\beta}}^{\mathcal{M}} + n_0 (\boldsymbol{\beta}_0^{\mathcal{M}})^T \boldsymbol{\beta}_0^{\mathcal{M}} \\ \tilde{\Sigma}_{\mathcal{M}} &= (n_0 \mathbb{I}_{p_{\mathcal{M}}} + \mathbf{X}_{\mathcal{M}}^T \mathbf{X}_{\mathcal{M}})^{-1} \\ \tilde{\boldsymbol{\beta}}^{\mathcal{M}} &= \tilde{\Sigma}_{\mathcal{M}} (n_0 \boldsymbol{\beta}_0^{\mathcal{M}} + \mathbf{X}_{\mathcal{M}}^T \mathbf{y}). \end{aligned}$$

Inserting this part into the integrand of the marginal probability results in

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \mathcal{M}) =& \int p(\mathbf{y}|\sigma^2, \mathbf{X}, \mathcal{M})\, p(\sigma^2)\, d\sigma^2 \\ =& (2\pi)^{-n/2} n_0^{p_{\mathcal{M}}/2} |\tilde{\Sigma}_{\mathcal{M}}|^{1/2} \left(\frac{\nu_0\sigma_0^2}{2}\right)^{\nu_0/2} \Gamma(\tfrac{\nu_0}{2})^{-1} \\ & \times \int \left(\frac{1}{\sigma^2}\right)^{(\nu_0+n)/2-1} \exp\left(-\frac{1}{\sigma^2}\frac{\nu_0\sigma_0^2 + \mathrm{SSR}_{n_0}^{\mathcal{M}}}{2}\right) d\sigma^2 \\ =& (2\pi)^{-n/2} n_0^{p_{\mathcal{M}}/2} |\tilde{\Sigma}_{\mathcal{M}}|^{1/2} \left(\frac{\nu_0\sigma_0^2}{2}\right)^{\nu_0/2} \Gamma(\tfrac{\nu_0}{2})^{-1} \times \frac{\Gamma((\nu_0+n)/2)}{([\nu_0\sigma_0^2 + \mathrm{SSR}_{n_0}^{\mathcal{M}}]/2)^{(\nu_0+n)/2}} \\ =& \pi^{-n/2} \frac{\Gamma((\nu_0+n)/2)}{\Gamma(\nu_0/2)} n_0^{p_{\mathcal{M}}/2} |\tilde{\Sigma}_{\mathcal{M}}|^{1/2} \frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_0\sigma_0^2 + \mathrm{SSR}_{n_0}^{\mathcal{M}})^{(\nu_0+n)/2}}, \end{aligned}$$

where the third equation follows after adding a normalizing constant to the unnormalised density in the integrand.

If n_0 is assigned a small number, then the prior for $\boldsymbol{\beta}$ is flat. Furthermore, for larger model dimensions $p_{\mathcal{M}}$, the likelihood decreases under this model. Therefore, models with more explanatory variables are penalized.

We now average over all possible models with a different set of regressors, so that the

distribution for new data $\tilde{y}$ is

$$p(\tilde{\mathbf{y}}|\mathbf{y}) = \sum_{l=1}^{2^{p-1}} p(\tilde{\mathbf{y}}|\mathcal{M}_l)\, p(\mathcal{M}_l|\mathbf{y}),$$

where $p(\tilde{\mathbf{y}}|\mathcal{M}_l)$ denotes the posterior distribution under model $\mathcal{M}_l$, and $p(\mathcal{M}_l|\mathbf{y})$ is the posterior probability of this model. The latter terms sum up to 1, and are the weights of the selected model.

3.3 Reference method

As a reference method, we investigate a simple method that pools all ensemble members and operates with their mean value only. The probabilistic forecasts results from a Gaussian distribution with parameters β and σ^2,

$$Y_i \sim \mathcal{N}\left(\beta + \bar{\mathbf{x}}_i, \sigma^2\right),$$

where $\bar{\mathbf{x}}_i = \sum_{j=1}^{p-1} x_{ij}$ denotes the mean of the ensemble members for $i = 1, \ldots, n$. The priors are set in the following way,

$$\begin{aligned} \beta|\sigma^2 &\sim \mathcal{N}\left(0, \frac{\sigma^2}{n_0}\right), \\ \frac{1}{\sigma^2} &\sim \Gamma\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right). \end{aligned}$$

Similar computations as in the previous sections lead to the following posteriors,

$$\begin{aligned} \frac{1}{\sigma^2} &\sim \Gamma\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + \sum^n (y_i - \bar{\mathbf{x}}_i)^2 - \frac{(\sum^n y_i - \bar{\mathbf{x}}_i)^2}{n_0+n}}{2}\right), \\ \beta|\sigma^2 &\sim \mathcal{N}\left(\frac{\sum^n y_i - \bar{\mathbf{x}}_i}{n_0 + n}, \frac{\sigma^2}{n_0 + n}\right). \end{aligned}$$

Similar as before, these posteriors provide us with a Monte Carlo sample from the joint posterior distribution $p(\sigma^2, \beta|\mathbf{y}, \mathbf{X})$.

The predictive distribution of a future value $\tilde{y}$ is conditioned on the observed data

used for estimating the parameters and has the form

$$\begin{aligned} p(\tilde{y}|\tilde{\mathbf{X}}, \mathbf{y}, \mathbf{X}) &= \int p(\tilde{y}, \boldsymbol{\beta}, \sigma^2|\tilde{\mathbf{X}}, \mathbf{y}, \mathbf{X})\, d\boldsymbol{\beta}\, d\sigma^2 \\ &= \int p(\tilde{y}|\tilde{\mathbf{X}}, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})\, d\boldsymbol{\beta}\, d\sigma^2. \end{aligned}$$

First, we obtain S samples from the joint posterior distribution $p(\sigma^2, \boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$, which are then used to get S samples of the Gaussian distribution $p(\tilde{y}|\tilde{\mathbf{X}}, \boldsymbol{\beta}, \sigma^2)$ with the corresponding posterior parameters $\boldsymbol{\beta}$ and σ^2. That way, we gain a sample from the predictive distribution.

3.4 Assessment of predictive performance

As we are now able to obtain forecasts for new data $\tilde{y}$ from the predictive distribution, the next step is to ascertain how precise our forecasts are and to assess their skill.

Our probabilistic forecasts should maximize the sharpness of our predictive distributions subject to calibration (Gneiting et al., 2007). Calibration associates the forecasting distributions and the verifying observations. To assess calibration, we will consider so-called Probability Integral Transform (PIT) histograms (Dawid, 1984; Gneiting et al., 2007). If y denotes the future value and F is the predictive cumulative distribution function, then the PIT value is $F(y)$. For calibrated forecasts, the histogram of the PIT values over the test set should resemble the uniform distribution. For the ensemble forecasts, a verification rank histogram is used to assess calibration. The histogram shows the relative frequencies of the ranks of the observed values when pooled within the ordered ensemble (Anderson, 1996; Hamill and Colucci, 1997). If the distribution of the ensemble predictions equals the distribution of the observations, a uniform histogram should emerge.

Moreover, we obtain the coverage of the prediction intervals to measure the forecast effectiveness. As our ensemble consists of eight ensemble members, the coverage of the 77.8%($= 100 \times (8-1)/(8+1)$) central prediction interval is examined, i. e. the proportion of the verifying observations that is within the 77.8% prediction interval. This corresponds to the proportion of observations that lie within the range of the raw ensemble forecast.

Sharpness relates to the spread of the forecast distributions, and to evaluate this, we provide the average width of the prediction intervals. Shorter intervals are preferred for more accurate predictions, subject to calibration.

Another option to measure the forecasting performance are proper scoring rules (Gneiting and Raftery, 2007). Deterministic forecasts can be evaluated by the Mean Absolute Error (MAE). For $k = 1, \ldots, K$ observations of the test data set, let the forecast $\hat{y}_k$ be deterministic, and the observation denoted by y_k, then the MAE is

$$\text{MAE} = \frac{1}{K} \sum_{k=1}^{K} |y_k - \hat{y}_k|.$$

Here, the determistic forecast is the median of the predictive distribution, i. e. the median of our sample of S values from the posterior predictive distribution.

A proper scoring rule for predictive density functions is the Continuous Ranked Probability Score (CRPS), which covers both calibration and sharpness (Matheson and Winkler, 1976; Hersbach, 2000; Gneiting and Raftery, 2007). If F denotes the predictive cumulative distribution function and y the observed value, then the CRPS is

$$\begin{aligned} \text{crps}(F, y) &= \int_{-\infty}^{\infty} [F(x) - \mathbb{I}(x \geq y)]^2 \, dx \\ &= \mathbb{E}|X - y| - \frac{1}{2}\mathbb{E}|X - X'|, \end{aligned}$$

where X and X' are independent variables with distribution F (Gneiting and Raftery, 2007). We divide our set of S forecasts in half with $S/2$ samples x_s and $S/2$ samples x'_s, where $s = 1, \ldots, S/2$. Then the mean of the difference between x_s and the verifying observation y is used as an approximation of $\mathbb{E}|X - y|$. The other term of the score is calculated likewise. That it,

$$\text{crps}(F, y) \approx \frac{1}{S/2} \sum_{s=1}^{S/2} |x_s - y| - \frac{1}{S} \sum_{s=1}^{S/2} |x_s - x'_s|.$$

For the test set we report the average CRPS, given by

$$\text{CRPS} = \frac{1}{K} \sum_{k=1}^{K} \text{crps}(F_k, y_k).$$

Both the MAE and the CRPS are negatively oriented and report values in the same units as the observation. When comparing the predictive perfomance of competing model forecasts, we thus aim to minimize the scores over the test set.

4 Case study

4.1 Description of the data

We apply the methods described in the previous chapter, to produce 48-h-ahead forecasts of surface temperature over the North American Pacific Northwest. Observations were available at 50 stations in the time period from January 1st, 2007 to December 31st, 2008, where the year 2008 was used as test period to evaluate the predictive performance of the various forecasting methods, bringing about a total of 14485 observations. The ensemble consists of the eight-member University of Washington Mesoscale Ensemble (UWME), see Eckel and Mass (2005). The forecasts are obtained with the fifth-generation Pennsylvania State University-National Center for Atmospheric Research Mesoscale Model with varying initial and boundary conditions, see Table 1. If there are less than eight ensemble members available or the verifying observations are absent, then this particular day is removed from the data set. Therefore, the training period often consisted of more days than the corresponding calendar time frame.

Furthermore, Table 2 shows that the ensemble member forecasts are highly correlated, so models including various ensemble members may be similarly skillfull.

Next, we have to decide upon the stations which should be included in the training set. One possibility is to select data from all stations for estimating the parameters, so

Table 1: The Centers that provide the initial and boundary conditions for the eight-member University of Washington Mesoscale Ensemble.

1	AVN, National Centers for Environmental Prediction
2	CMCG, Canadian Meteorological Centre
3	ETA, National Centers for Environmental Prediction
4	GASP, Australian Bureau of Meteorology
5	JMA, Japan Meteorological Agency
6	NGPS, Fleet Numerical Meteorological & Oceanographic Center
7	TCWB, Taiwan Central Weather Bureau
8	UKMO, Met Office

Table 2: Correlations between the ensemble temperature forecasts.

	AVN	CMCG	ETA	GASP	JMA	NGPS	TCWB	UKMO
AVN	1.00	0.93	0.92	0.92	0.91	0.92	0.93	0.92
CMCG	0.93	1.00	0.92	0.92	0.93	0.91	0.92	0.93
ETA	0.92	0.92	1.00	0.92	0.92	0.92	0.91	0.92
GASP	0.92	0.92	0.92	1.00	0.92	0.92	0.92	0.91
JMA	0.91	0.93	0.92	0.92	1.00	0.92	0.92	0.93
NGPS	0.92	0.91	0.92	0.92	0.92	1.00	0.92	0.92
TCWB	0.93	0.92	0.91	0.92	0.92	0.92	1.00	0.92
UKMO	0.92	0.93	0.92	0.91	0.93	0.92	0.92	1.00

that on each day, the forecasts at every station are computed using the same, regionally estimated parameters. This method is referred to as the regional method, where we have 291 test days in total. Our study also considers a local, station-specific approach, where the parameters are estimated at each station separately by using only the observations of this particular station for training. We refer to this as the local model.

4.2 Choice of prior distribution

In Chapter 3, we discussed three possible choices for the prior distribution of our BEMOS method. In the regional setting, the conjugate prior distribution in Chapter 3.1.2 and the g-prior in Chapter 3.1.3 yield comparable predictive performance. However, the prior distribution in Chapter 3.1.2 requires a Markov chain Monte Carlo approximation with a Gibbs sampler to obtain samples from the posterior predictive distributions. This method is thus computationally extremely intensive and in that respect not comparable to alternative methods such as EMOS (Gneiting et al., 2005) and BMA (Raftery et al., 2005).

In the local setting, the weakly informative prior in Chapter 3.1.4 yields significantly better predictive performance than the g-prior discussed in Chapter 3.1.3. For these reasons, we focus on the weakly informative prior in the following and especially discuss the choice of the prior parameter n_0.

4.3 Choice of training period

To find a forecast on day j, we choose a training period of n days, consisting of the days $j-(n+1), \ldots, j-2$, as we are aiming to predict the observation two days ahead. We have

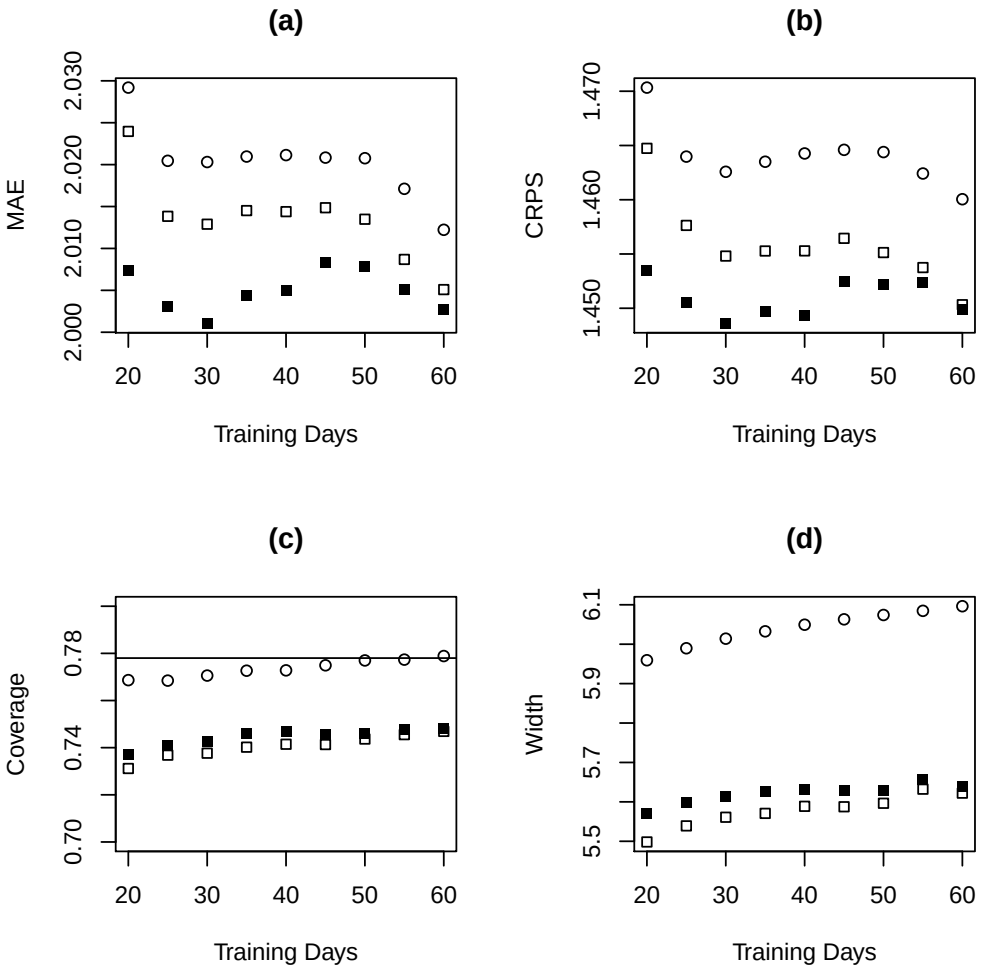

Figure 1: Different length of the training period for temperature forecasts under the regional method: (a) Mean Absolute Error, (b) Continuous Ranked Probability Score, (c) Coverage of the 77.8% intervals, (d) Width of the 77.8% prediction intervals; (○) BEMOS ($n_0 = 0.01$), (□) EMOS, (■) EMOS$^+$.

to determine an appropriate choice of n. For this, we examine periods of $20, 25, \ldots, 55, 60$ days in the regional case. We consider the models EMOS, EMOS$^+$ and BEMOS with the weakly informative prior discussed in Chapter 3.1.4. Here, n_0 is set to 0.01, which corresponds to a high variability of the coefficients $\boldsymbol{\beta}$ and a flat prior distribution. As one can see in Figure 1, the MAE and the mean CRPS decrease at first, increase after a training time of 30 days, and decrease again. We also see that the BEMOS method covers the 77.8% prediction interval quite well for all training days, whereas the EMOS

methods show a worse coverage. Finally, the last graph shows the average width of the prediction intervals which increases if more days are included in the training period. The intervals of the BEMOS method are slightly wider than the intervals of the competing models. On the whole, each graph shows that all three methods are not sensitive to changes in the length of the training period as the values are nearly constant. Based on these results, we use a sliding window of 30 days for the regional methods.

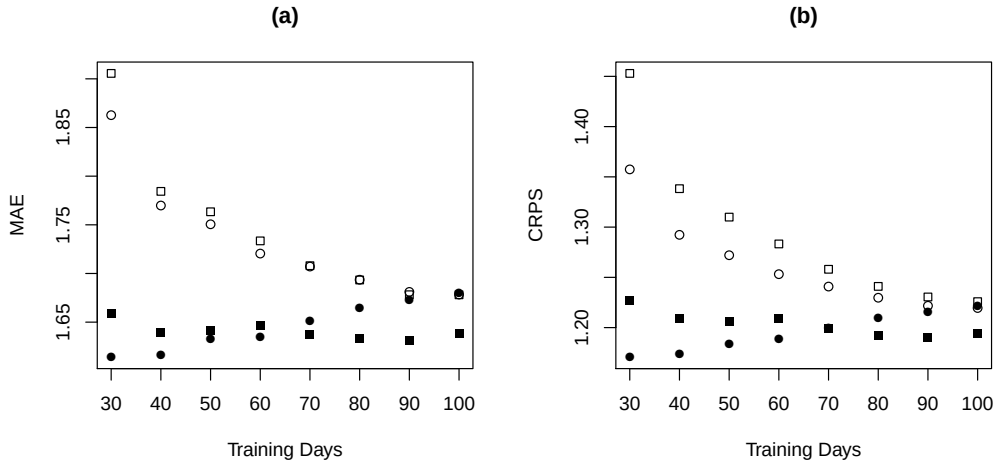

Figure 2: Different length of the training period for temperature forecasts of the local model at 20 randomly chosen stations: (a) Mean Absolute Error; (b) Continuous Ranked Probability Score; (○) BEMOS ($n_0 = 0.01$), (●) BEMOS ($n_0 = 200$), (□) EMOS, (■) EMOS$^+$.

A similar procedure is performed for the local methods. The training set is here composed of data from a single station and in this setting, the length of the training period has a greater impact on the predictive performance. We consider $30, \ldots, 100$ days, but due to computational reasons we use only 20 of the 50 stations which are chosen randomly. Figure 2 shows that the MAE and mean CRPS of the BEMOS model with parameter $n_0 = 0.01$ decreases with longer periods and that there is now a great difference between the EMOS and EMOS$^+$ method. Like BEMOS, EMOS improves as the training period increases, whereas in comparison, EMOS$^+$ yields overall better results which do not vary very much. We observed that there were some days where the EMOS$^+$ method dropped every ensemble member, so the estimate was just the intercept. And yet, this estimation leads to smaller values of the MAE and the CRPS. In comparison to the BEMOS model with parameter $n_0 = 0.01$, if we set $n_0 = 200$, we

see that increased size of the training set has an opposite effect and the values vary less strongly. As for the regional method, the length of the training period is here 30 days.

4.4 Suitable parameters

For the parameters of the weakly informative prior, we set $\sigma_0^2 = 1$ and $\nu_0 = 1$. The prior distribution for the coefficients $\boldsymbol{\beta}$ has a mean of
$\boldsymbol{\beta}_0 = (0, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8)$, so that each ensemble member is a priori equally likely. We consider several values for n_0, where a small number indicates a large variability of $\boldsymbol{\beta}$ around $\boldsymbol{\beta}_0$. Table 3 lists the resulting MAE and mean CRPS values for the regional and the local BEMOS methods over the test set for different choices of n_0. The scores are better for large values of n_0, especially for the local method. Hence, our model is very sensitive to the choice of the parameter n_0.

Table 3: MAE and mean CRPS over the test set for the BEMOS model with changing parameter n_0.

	Regional model		Local model	
n_0	MAE	CRPS	MAE	CRPS
0.01	2.02	1.46	1.90	1.38
0.1	2.01	1.46	–	–
1	2.00	1.45	1.88	1.37
100	1.99	1.45	1.66	1.20
200	1.99	1.44	1.66	1.20
500	1.98	1.44	1.66	1.20
1000	1.98	1.44	–	–

4.5 Temperature forecasts

Finally, we compare the Bayesian BEMOS approach to the performance of both EMOS and EMOS$^+$.

First, we examine the estimated regional regression coefficients of BEMOS and EMOS$^+$ shown in Figure 3, where the parameters are fitted using training data of all 50 stations in a 30-day training period, resulting in one common parameter set for all stations. All along, the BEMOS intercept represented in plot (a) is about zero, whereas the EMOS$^+$ technique yields values that range from about -20 to 45. The regression coefficients of the ensemble members shown in plots (b)-(f) vary over the course of time and are quite

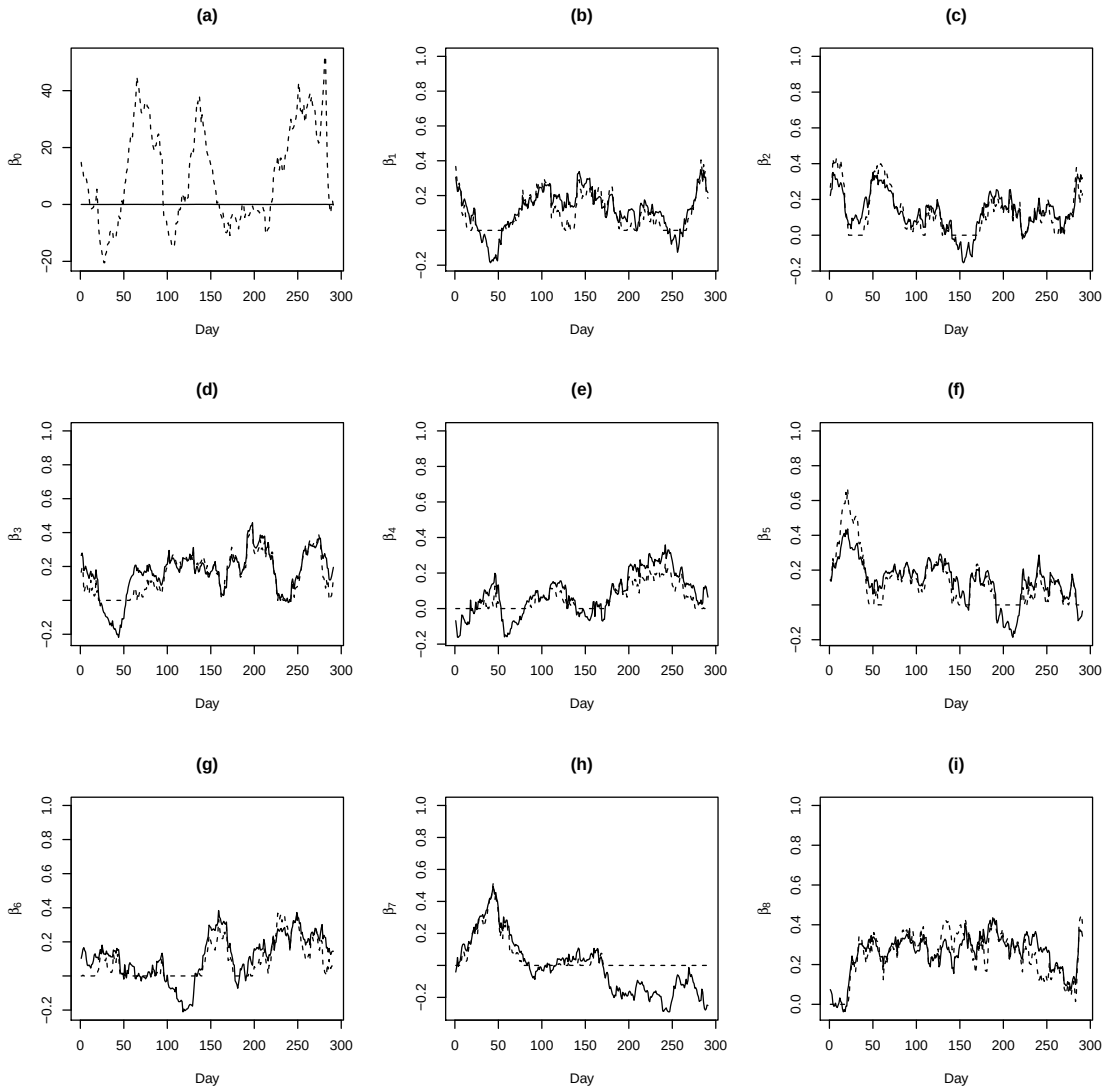

Figure 3: Regional EMOS$^+$ and BEMOS coefficients for temperature forecasts over the Pacific Northwest for the 291 available days in the year 2008: (a) intercept, (b)-(i) ensemble member weights; (—) BEMOS ($n_0 = 500$), (---) EMOS$^+$.

similar except for the cases when the BEMOS coefficients are negative. In this period of time, EMOS$^+$ drops the ensemble member from the model.

In the local technique each station has a different set of parameter estimates. In Figure 4, the estimated intercept and coefficients at the station Sea-Tac Airport are reported. Again, the intercept of the BEMOS method is close to zero, while the EMOS$^+$ intercept varies strongly over time. In fact, there are periods at some stations where EMOS$^+$

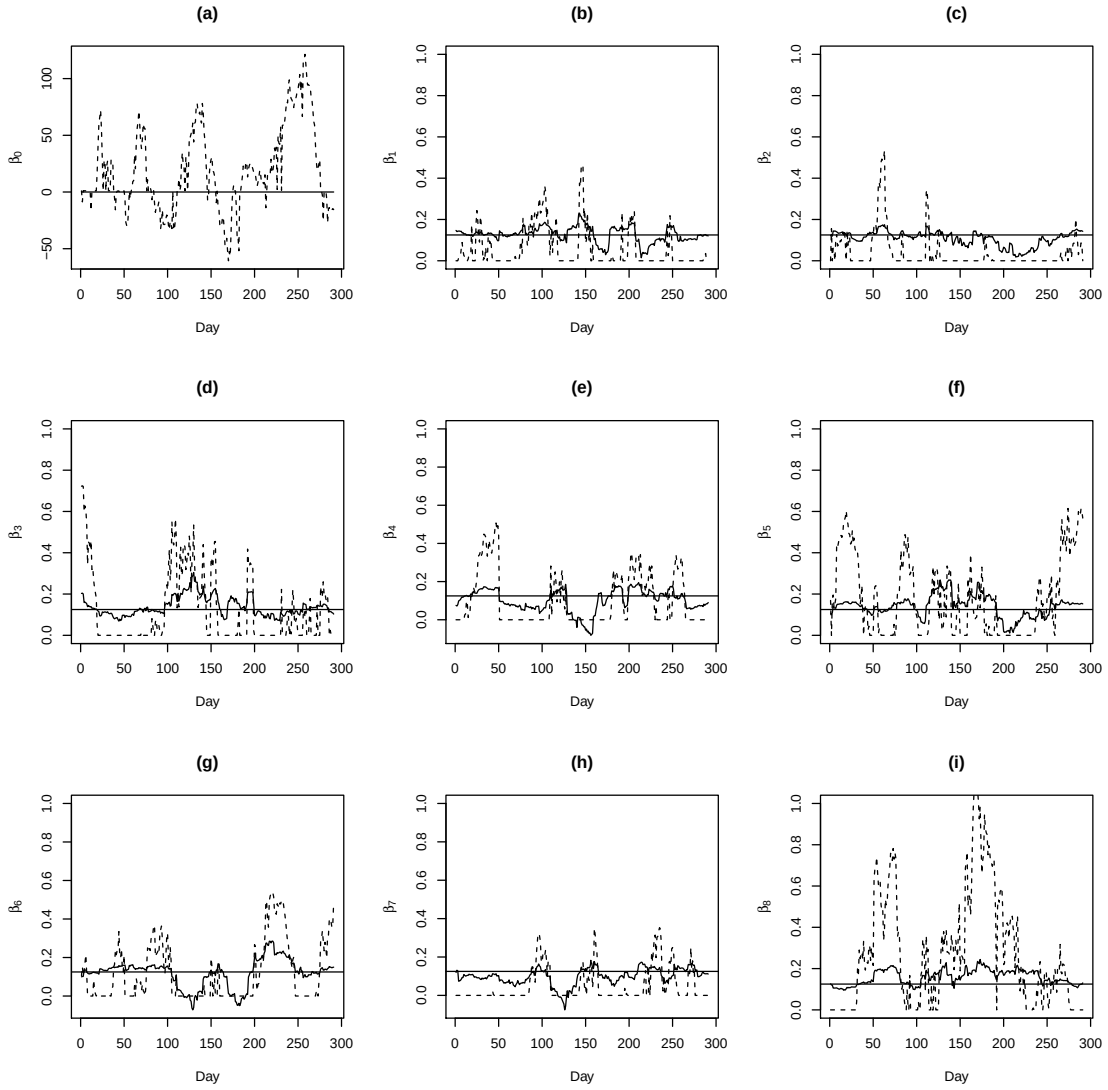

Figure 4: Local EMOS$^+$ and BEMOS coefficients for temperature forecasts at the station Sea-Tac Airport for the 291 available days in the year 2008: (a) intercept, (b)-(i) ensemble member weights; (—) BEMOS ($n_0 = 200$), (- - -) EMOS$^+$.

drops every ensemble member and estimates the forecast using only the fitted intercept. Furthermore, its coefficients fluctuate strongly over the period, and it often drops several members on a given day. On the other hand, all BEMOS coefficients vary around the value 1/8 (represented by the horizontal line) for the most part, so that this method shrinks the forecast to be close to the ensemble mean.

Table 4 summarizes the results of the measures of predictive performance of the com-

Table 4: Prediction performance of the regional models: Scores, coverage and the average width of the 50%, 77.8% and 90% prediction intervals for the raw and bias-corrected ensemble, and the EMOS and BEMOS methods, the last being the model, where every possible combination of the ensemble members are included.

	Scores		Prediction Interval					
			77.8%		50%		90%	
Forecasts	MAE	CRPS	Cov.	Width	Cov.	Width	Cov.	Width
Ens.raw	2.08	1.69	0.39	2.83	–	–	–	–
Ens.bc	2.03	1.67	0.39	2.63	–	–	–	–
EMOS	2.01	1.45	0.74	5.56	0.49	3.07	0.86	7.49
EMOS$^+$	2.00	1.45	0.74	5.61	0.49	3.10	0.86	7.57
BEMOS$_{\text{sel}}$($n_0 = 0.01$)	2.02	1.51	0.61	4.25	0.39	2.35	0.73	5.73
BEMOS ($n_0 = 0.01$)	2.02	1.46	0.77	6.01	0.52	3.32	0.88	8.11
BEMOS ($n_0 = 1$)	2.00	1.45	**0.78**	6.05	0.52	3.34	0.88	8.15
BEMOS ($n_0 = 500$)	**1.98**	**1.44**	**0.78**	6.09	0.53	3.37	0.89	8.21
BEMOS$_{\text{ref}}$($n_0 = 500$)	1.99	**1.44**	0.79	6.27	0.54	3.46	0.89	8.45

peting regional models, where we compare values of the MAE and the mean CRPS, and examine the coverage and mean width of the prediction intervals. The forecasts of the bias-corrected ensemble are estimated using linear least squares regression on the ensemble members. The prediction intervals of the raw and bias-corrected ensemble are very sharp but have a poor coverage and have by far the worst scores. The method BEMOS$_{\text{sel}}$, where each possible model is considered, also performs badly, while being computationally intensive. We continue to work with the full Bayesian model including all eight ensemble members. As we have already seen above, BEMOS improves with higher values of n_0, and since the choice of $n_0 = 100, 200, 500, 1000$ result in nearly the same scores, we select the method with parameter $n_0 = 500$. Moreover, we settle for the reference model with $n_0 = 500$, where the parameters are estimated considering the ensemble mean only. The predictive PDFs of EMOS and EMOS$^+$ are underdispersive, whereas the predictive distributions of the BEMOS methods showed an accurate coverage but, in comparison, sligthly wider intervals.

Figure 5 shows the verification rank histograms for the full raw and bias-corrected data set, and PIT histograms for the methods EMOS, EMOS$^+$ and BEMOS. The histograms of the raw and the bias-corrected ensemble are U-shaped, which indicates underdispersion, so that the observations incline to be smaller (greater) than the minimum (maximum) of the ensemble range. As the histograms (c)-(h) for the other techniques

show, EMOS and BEMOS are better calibrated with close to uniform PIT histograms.

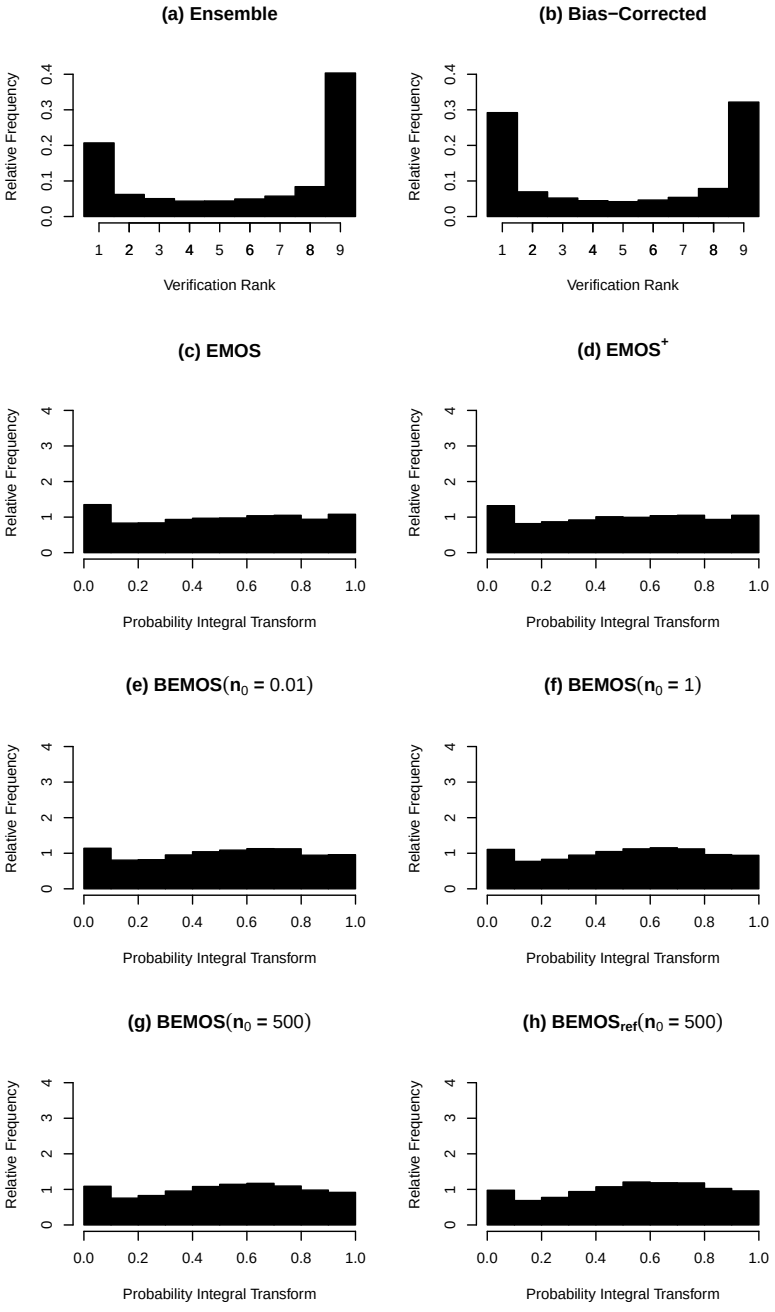

Figure 5: Verification rank histograms for the (a) raw and (b) bias-corrected ensemble, and (c)-(h) PIT histograms for the competing temperature forecasts of the regional models over all available data in 2008.

The probabilistic forecasts of the local models are assessed similarly, as summarized in Table 5. The technique EMOS$^+$ have a much better MAE and mean CRPS than EMOS. However, both methods are uncalibrated. The performance of the BEMOS methods with small n_0-values is also insufficient. Meanwhile, BEMOS ($n_0 = 200$) exhibits the best scores and a decent coverage of the considered prediction intervals (similar results are obtained with $n_0 = 100, 500$ and these are thus omitted). A summary of the predictive performance at the Sea-Tac station is furthermore given in Table 6.

Table 5: Prediction Performance of the local models over all stations and all available days in 2008: Scores, coverage and the average width of the 50%, 77.8% and 90% prediction intervals for the raw and bias-corrected ensemble, and the EMOS and BEMOS methods.

	Scores		Prediction Interval					
			77.8%		50%		90%	
Forecasts	MAE	CRPS	Cov.	Width	Cov.	Width	Cov.	Width
Ens.raw	2.08	1.69	0.39	2.83	–	–	–	–
Ens.bc	1.69	1.36	0.47	2.66	–	–	–	–
EMOS	1.93	1.47	0.53	3.33	0.32	1.84	0.67	4.49
EMOS$^+$	1.68	1.25	0.63	3.75	0.40	2.07	0.77	5.06
BEMOS ($n_0 = 0.01$)	1.90	1.38	0.69	4.67	0.43	2.56	0.82	6.34
BEMOS ($n_0 = 1$)	1.88	1.37	0.69	4.70	0.44	2.58	0.82	6.40
BEMOS ($n_0 = 200$)	**1.66**	**1.20**	0.76	4.85	0.51	2.65	0.87	6.60
BEMOS$_{\text{ref}}$($n_0 = 0.01$)	**1.66**	1.21	0.77	4.90	0.51	2.68	0.88	6.66
BEMOS$_{\text{ref}}$($n_0 = 200$)	1.98	1.40	**0.78**	6.03	0.49	3.31	0.90	8.21

Table 6: Prediction Performance of the local methods at Sea-Tac Airport in 2008: Scores, coverage and the average width of the 50%, 77.8% and 90% prediction intervals for the raw and bias-corrected ensemble, and the EMOS and BEMOS methods.

	Scores		Prediction Interval					
			77.8%		50%		90%	
Forecasts	MAE	CRPS	Cov.	Width	Cov.	Width	Cov.	Width
Ens.raw	1.64	1.27	0.47	3.04	–	–	–	–
Ens.bc	1.49	1.15	0.43	2.71	–	–	–	–
EMOS	1.82	1.39	0.51	3.13	0.34	1.73	0.64	4.22
EMOS$^+$	1.50	1.11	0.62	3.45	0.43	1.90	0.78	4.64
BEMOS ($n_0 = 200$)	1.50	1.07	0.76	4.38	0.46	2.40	0.89	5.96

Figure 6 displays the histograms of the local methods. However, the histograms for the EMOS, EMOS$^+$ and BEMOS ($n_0 = 0.01, 1$) forecasts are also slightly U-shaped, whereas the histogram for the BEMOS ($n_0 = 200$) forcasts are close to being uniformly distributed.

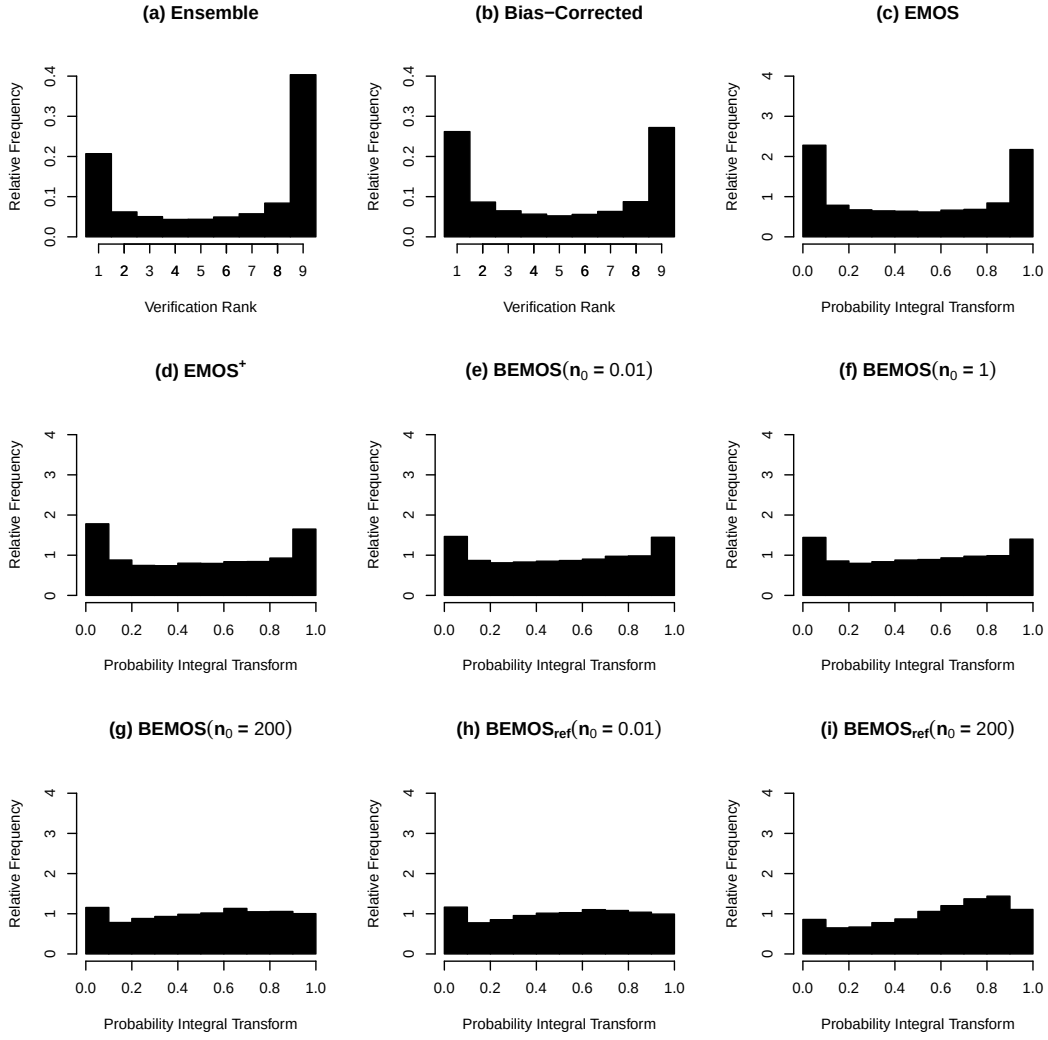

Figure 6: Verification rank histograms for the (a) raw and (b) bias-corrected ensemble, and (c)-(i) PIT histograms for the competing temperature forecasts of the local models over all stations and available data in 2008.

5 Conclusion

We propose a fully Bayesian method to produce probabilistic forecasts for temperature that is based on the EMOS method of Gneiting et al. (2005). The model consists of Gaussian predictive distributions with a mean that is a weighted average of deterministic ensemble forecasts, corrected for bias. Several choices of the multivariate normal prior distribution for the regression coefficients are considered, where a weakly informative prior performs best. The variance of this prior depends on a parameter n_0 which regulates the spread of the prior with a small value indicating a flat prior distribution. The variance of our predictive distribution is assigned an inverse-gamma distribution. After observations of a training data set are included in the model, the parameters are updated and posterior distributions are formed. Both priors are conjugate resulting in likewise multivariate normal distribution and an inverse-gamma distribution, respectively. To assess the model skill, predictions for a test data set are obtained.

The year 2008 is used as a test period to produce 48-h-ahead forecasts of surface temperature over the North American Pacific Northwest. Our data set consists of the University of Washington Mesoscale Ensemble (UWME) with eight ensemble members. When applying our model over the test period in the year 2008, we see that it strongly depends on the choice of the parameter n_0, and assigning large values to n_0 results in the lowest values of the MAE and the mean CRPS. Moreover, in comparison to the competing EMOS and EMOS$^+$ methods, we get a significantly better calibration. While the EMOS$^+$ model drops some ensemble members, BEMOS develops a shrinkage towards the ensemble mean with some alterations. Our reference model in the global approach, where only the ensemble mean is used as a covariate, behaves similarly well. In contrast to this result, the reference model in the local case performs better with a flat prior (when n_0 is small). There might be room for improvement if an automatic way of determining the parameter n_0 can be found, for instance in form of an appropriate prior distribution.

On the whole, our prediction model leads to an improvement when assessing predictive perfomance, where we get an accurate coverage, i. e. calibrated forecasts, and good values in terms of the scoring rules MAE and CRPS.

Bibliography

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate 9*, 1518–1530.

Britannica (2012). Weather forecasting — Britannica - the online encyclopedia. `http://www.britannica.com/EBchecked/topic/638321/weather-forecasting/49626/Numerical-weather-prediction-NWP-models?anchor=ref293532` [Online; accessed 25-March-2012].

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society 147A*, 278–292.

Eckel, F. A. and C. F. Mass (2005). Aspects of effective mesoscale, short-range ensemble forecasting. *Weather and Forecasting*, 328–350.

Glahn, H. R. and D. A. Lowry (1972). The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*.

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

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

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. *Monthly Weather Review 133*, 1098–1117.

Hamill, T. M. and S. J. Colucci (1997). Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review 125*, 1312–1327.

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

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer.

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

Narzo, A. F. D. and D. Cocchi (2010). A Bayesian hierarchical approach to ensemble weather forecasting. *Journal of the Royal Statistical Society 59C*(3), 405–422.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review 133*, 1155–1173.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Volume 6, pp. 233–243. North Holland. Studies in Bayesian Econometrics and Statistics.

Erklärung

Hiermit versichere ich, dass ich meine Arbeit selbstständig unter Anleitung verfasst habe, dass ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, und dass ich alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entlehnt sind, durch die Angabe der Quellen als Entlehnung kenntlich gemacht habe.

Datum Unterschrift