132

ensembleBMA: AN R PACKAGE FOR PROBABILISTIC WEATHER FORECASTING

Chris Fraley*, Adrian Raftery
University of Washington, Seattle, WA USA
Tilmann Gneiting
University of Heidelberg, Heidelberg, Germany
McLean Sloughter
Seattle University, Seattle, WA USA

Abstract. This article describes an R package for probabilistic weather forecasting, `ensembleBMA`, which offers ensemble postprocessing via Bayesian model averaging (BMA). BMA forecasting models use mixture distributions, in which each component corresponds to an ensemble member, and the form of the component distribution depends on the weather parameter (temperature, quantitative precipitation or wind speed). The model parameters are estimated from training data. The package includes functions for evaluating the predictive performance, in addition to model fitting and forecasting.

1 INTRODUCTION

Over the past two decades, weather forecasting has experienced a paradigm shift towards probabilistic forecasts, which take the form of probability distributions over future weather quantities and events. Probabilistic forecasts allow for optimal decision making for many purposes, including air traffic control, ship routing, agriculture, electricity generation and weather-risk finance.

Up to the early 1990s, most weather forecasting was deterministic, meaning that only one "best" forecast was produced by a numerical model. The recent advent of ensemble prediction systems marks a radical change. An ensemble forecast consists of multiple numerical forecasts, each computed in a different way. Statistical postprocessing is then used to convert the ensemble into calibrated and sharp probabilistic forecasts (Gneiting and Raftery 2005).

The `ensembleBMA` package (Fraley et al. 2006) offers statistical postprocessing of forecast ensembles via Bayesian model averaging (BMA). It provides functions for model fitting and forecasting with ensemble data that may include missing and/or exchangeable members. The modeling functions estimate BMA parameters from training data via the EM algorithm. Currently available options are normal mixture models (appropriate for temperature or pressure), mixtures of gamma distributions (appropriate for wind speed), and Bernoulli-gamma mixtures with a point mass at zero (appropriate for quantitative precipitation). Functions for verification that assess the predictive performance are also available.

The BMA approach to the postprocessing of ensemble forecasts was introduced by Raftery et al. (2005) and has been developed in Berrocal et al. (2007), Sloughter et al. (2007), Wilson et al. (2007), Fraley et al. (2010) and Sloughter et al. (2010). Detail on verification procedures can be found in Gneiting et al. (2007) and Gneiting and Raftery (2007).

2 ensembleData OBJECTS

Ensemble forecasting data for weather typically includes some or all of the following information:

- ensemble member forecasts
- initial date
- valid date
- forecast hour (prediction horizon)
- location (latitude, longitude, elevation)
- station and network identification

The "initial date" is the day and time at which initial conditions are provided to the numerical weather prediction model, to run forward the partial differential equations that produce the members of the forecast ensemble. The "forecast hour" is the prediction horizon or time between initial and valid dates. The ensemble member forecasts then are valid for the

* *Corresponding author address:* Dept of Statistics, Box 354322, U of Washington, Seattle, WA 98915-4322

hour and day that correspond to the "forecast hour" ahead of the "initial date". In all the examples and illustrations in this article, the prediction horizon is 48 hours.

For use with the ensembleBMA package, data must be organized into an ensembleData object that minimally includes the ensemble member forecasts. For model fitting and verification, the corresponding weather observations are also needed. Several of the model fitting functions can produce forecasting models over a sequence of dates, provided that the ensembleData are for a single prediction horizon. Attributes such as station and network identification, and latitude and longitude, may be useful for plotting and/or analysis but are not currently used in any of the modeling functions. The ensembleData object facilitates preservation of the data as a unit for use in processing by the package functions.

Here we illustrate the creation of an ensembleData object called srftData that corresponds to the srft data set of surface temperature (Berrocal et al. 2007):

```
data(srft)
members <- c("CMCG", "ETA", "GASP", "GFS",
             "JMA", "NGPS", "TCWB", "UKMO")
srftData <-
 ensembleData(forecasts = srft[,members],
 dates = srft$date, observations = srft$obs,
 latitude = srft$lat, longitude = srft$lon,
 forecastHour = 48)
```

The "dates" specification in an ensembleData object refers to the valid dates for the forecasts.

Specifying exchangeable members. Forecast ensembles may contain members that can be considered exchangeable (arising from the same generating mechanism, such as random perturbations of a given set of initial conditions), and for which the BMA parameters, such as weights and bias correction coefficients, should be the same. In ensembleBMA, exchangeability is specified by supplying a vector that represents the grouping of the ensemble members. The non-exchangeable groups consist of singleton members, while exchangeable members belong to the same group. See Fraley et al. (2010) for a detailed discussion.

Specifying dates. Functions that rely on the chron package (Hornik 1999) are provided for converting to and from Julian dates. The functions check for proper format (YYYYMMDD or YYYYMMDDHH).

3 BMA FORECASTING

BMA generates full predictive probability density functions (PDFs) for future weather quantities. Examples of BMA predictive PDFs for temperature and precipitation are shown in Figure 1.

Surface temperature example. As an example, we fit a BMA normal mixture model for forecasts of surface temperature valid January 31, 2004, using the srft training data. The ensembleData object srftData created in the previous section is used to fit the predictive model, with a rolling training period of 25 days, excluding the two most recent days because of the 48 hour prediction horizon.

One of several options is to use the function ensembleBMA with the valid date(s) of interest as input to obtain the associated BMA fit(s):

```
srftFit <-
  ensembleBMA( srftData, dates = "2004013100",
          model = "normal", trainingDays = 25)
```

When no dates are specified, a model fit is produced for each date for which there are sufficient training data for the desired rolling training period.

The BMA predictive PDFs can be plotted as follows, with Figure 1 showing an example:

```
plot( srftFit, srftData, dates = "2004013100")
```

This steps through each location on the given dates, plotting the corresponding BMA PDFs.

Alternatively, the modeling process for a single date can be separated into two steps: first extracting the training data, and then fitting the model directly using the fitBMA function. See Fraley et al. (2007) for an example. A limitation of this approach is that date information is not automatically retained.

Forecasting is often done on grids that cover an area of interest, rather than at station locations. The dataset srftGrid provides ensemble forecasts of surface temperature initialized on January 29, 2004 and valid for January 31, 2004 at grid locations in the same region as that of the srft stations.

Quantiles of the BMA predictive PDFs at the grid locations can be obtained with the function quantileForecast:

```
srftGridForc <- quantileForecast( srftFit,
  srftGridData, quantiles = c( .1, .5, .9))
```

Here srftGridData is an ensembleData object created from srftGrid, and srftFit provides a fore-

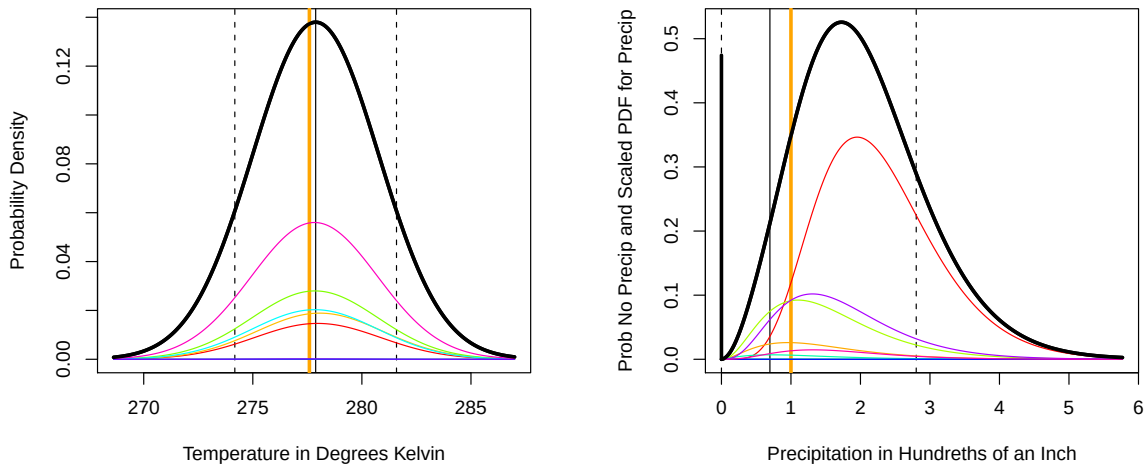

Figure 1: BMA predictive distributions for temperature (in degrees Kelvin) valid January 31, 2004 (left) and for precipitation (in hundredths of an inch) valid January 15, 2003 (right), at Port Angeles, Washington at 4PM local time, based on the eight-member University of Washington Mesoscale Ensemble (Grimit and Mass 2002; Eckel and Mass 2005). The thick black curve is the BMA PDF, while the colored curves are the weighted PDFs of the constituent ensemble members. The thin vertical black line is the median of the BMA PDF (occurs at or near the mode in the temperature plot), and the dashed vertical lines represent the 10th and 90th percentiles. The orange vertical line is at the verifying observation. In the precipitation plot (right), the thick vertical black line at zero shows the point mass probability of no precipitation (47%). The densities for positive precipitation amounts have been rescaled, so that the maximum of the thick black BMA PDF agrees with the probability of precipitation (53%).

casting model for the correspondong date.[1] The forecast probability of temperatures below freezing at the grid locations can be computed with the `cdf` function, which evaluates the BMA cumulative distribution function:

```
probFreeze <- cdf( srftFit, srftGridData,
        date = "2004013100", value = 273.15)
```

In the `srft` and `srftGrid` datasets, temperature is recorded in degrees Kelvin (K), so freezing occurs at 273.15 K.

These results can be displayed as image plots using the `plotProbcast` function, as shown in Figure 2, in which darker shades represent higher probabilities. The plots are made by binning values onto a plotting grid, which is the default in `plotProbcast`. Loading the `fields` (Furrer et al. 2001) and `maps` (Brownrigg and Minka 2003) libraries enables display of the country and state outlines, as well as a legend.

[1]The package implements the original BMA method of Raftery et al. (2005) and Sloughter et al. (2007), in which there is a single, constant bias correction term over the whole domain. Model biases are likely to differ by location, and there are newer methods that account for this (Gel 2007; Mass et al. 2008; Kleiber et al. 2011).

Precipitation example. The `prcpFit` dataset consist of the fitted BMA parameters for 48 hour ahead forecasts of daily precipitation accumulation (in hundredths of an inch) over the U.S. Pacific Northwest from December 12, 2002 through March 31, 2005, as described by Sloughter et al. (2007). The fitted models are Bernoulli-gamma mixtures with a point mass at zero that apply to the cube root transformation of the ensemble forecasts and observed data. A rolling training period of 30 days is used. The dataset used to obtain `prcpFit` is not included in the package on account of its size. However, the corresponding `ensembleData` object can be constructed in the same way as illustrated for the surface temperature data, and the modeling process also is analogous, except that the `"gamma0"` model for quantitative precipitation is used in lieu of the `"normal"` model.

The `prcpGrid` dataset contains gridded ensemble forecasts of daily precipitation accumulation in the same region as that of `prcpFit` initialized January 13, 2003 and valid January 15, 2003. The BMA median and upper bound (90th percentile) forecasts can be obtained and plotted as follows:

```
data(prcpFit)

prcpGridForc <- quantileForecast(
```

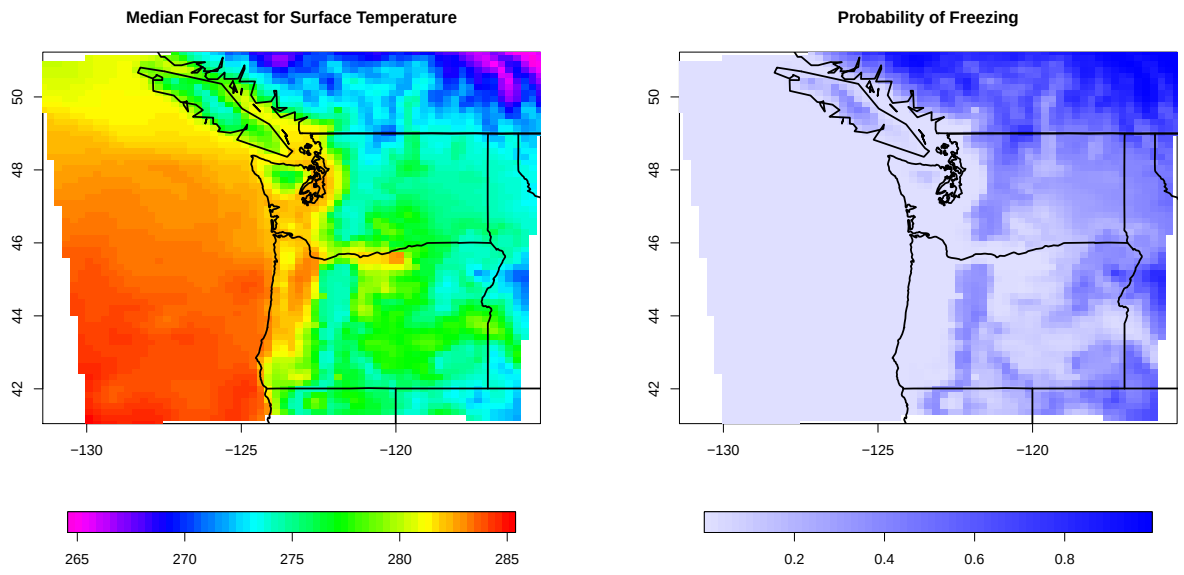

Figure 2: Image plots of the BMA median forecast for surface temperature and BMA probability of freezing over the Pacific Northwest valid January 31, 2004.

```
  prcpFit, prcpGridData, date = "20030115",
  q = c(0.5, 0.9))
```

Here `prcpGridData` is an `ensembleData` object created from the `prcpGrid` dataset. The 90th percentile plot is shown in Figure 3. The probability of precipitation and the probability of precipitation above .25 inches can be obtained as follows:

```
probPrecip <- 1 - cdf( prcpFit, prcpGridData,
  date = "20030115", values = c(0, 25))
```

The plot for the BMA forecast probability of precipitation accumulation exceeding .25 inches is also shown in Figure 3.

4 VERIFICATION

The `ensembleBMA` functions for verification can be used whenever observed weather conditions are available. Included are functions to compute verification rank and probability integral transform histograms, the mean absolute error, continuous ranked probability score, and Brier score.

Mean absolute error, continuous ranked probability score, and Brier score. In the previous section, we obtained a gridded BMA forecast of surface temperature valid January 31, 2004 from the `srft` data set. To obtain forecasts at station locations, we apply the function `quantileForecast` to the model fit `srftFit`:

```
srftForc <- quantileForecast( srftFit,
  srftData, quantiles = c( .1, .5, .9))
```

The BMA quantile forecasts can be plotted together with the observed weather conditions using the function `plotProbcast` as shown in Figure 4. Here the R core function `loess` was used to interpolate from the station locations to a grid for surface plotting. It is also possible to request image or perspective plots, or contour plots.

The mean absolute error (MAE) and mean continuous ranked probability score (CRPS; e.g., Gneiting and Raftery 2007) can be computed with the functions `CRPS` and `MAE`:

```
CRPS( srftFit, srftData)
# ensemble      BMA
# 1.945544 1.490496

MAE( srftFit, srftData)
# ensemble      BMA
# 2.164045 2.042603
```

The function `MAE` computes the mean absolute difference between the ensemble or BMA median forecast[2] and the observation. The BMA CRPS is obtained via Monte Carlo simulation and thus may vary slightly in replications. Here we compute these measures from forecasts valid on a single date; more typically, the CRPS and MAE would be computed from

[2]Raftery et al. (2005) employ the BMA predictive mean rather than the predictive median.

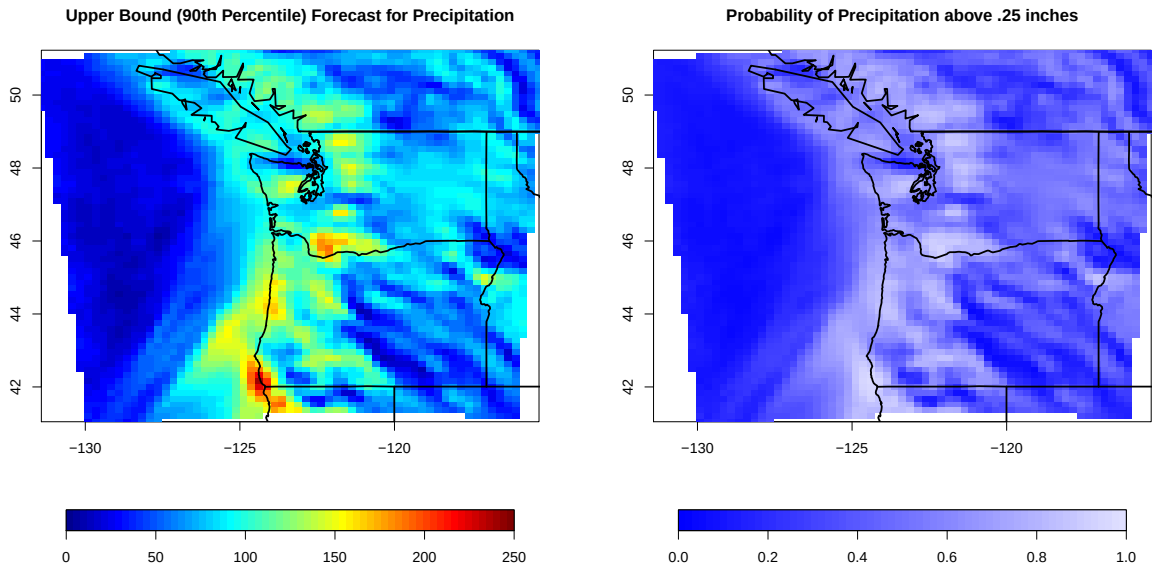

Figure 3: Image plots of the BMA upper bound (90th percentile) forecast of precipitation accumulation (in hundredths of an inch), and the BMA probability of precipitation exceeding .25 inches over the Pacific Northwest valid January 15, 2003.

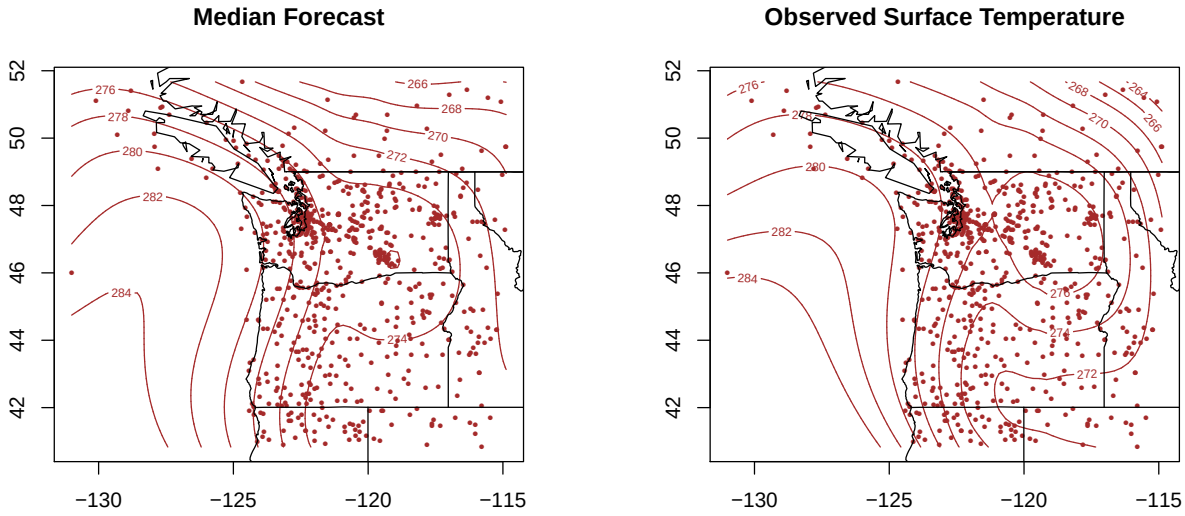

Figure 4: Contour plots of the BMA median forecast of surface temperature and verifying observations at station locations in the Pacific Northwest valid January 31, 2004 (`srft` dataset). The plots use `loess` fits to the forecasts and observations at the station locations, which are interpolated to a plotting grid. The dots represent the 715 observation sites.

a range of dates and the corresponding predictive models.

Brier scores (see e.g. Joliffe and Stephenson 2003 or Gneiting and Raftery 2007) for probability forecasts of the binary event of exceeding an arbitrary precipitation threshold can be computed with the function `brierScore`.

Assessing calibration. Calibration refers to the statistical consistency between the predictive distributions and the observations (Gneiting et al. 2007). The *verification rank histogram* is used to assess calibration for an ensemble forecast, while the *probability integral transform* (PIT) histogram assesses calibration for predictive PDFs, such as the BMA forecast distributions.

The verification rank histogram plots the rank of each observation relative to the combined set of the ensemble members and the observation. Thus, it equals one plus the number of ensemble members that are smaller than the observation. The histogram allows for the visual assessment of the calibration of the ensemble forecast (Hamill 2001). If the observation and the ensemble members are exchangeable, all ranks are equally likely, and so deviations from uniformity suggest departures from calibration. We illustrate this with the `srft` dataset, starting at Januray 30, 2004:

```
use <- ensembleValidDates(srftData) >=
              "2004013000"

srftVerifRank <- verifRank(
  ensembleForecasts(srftData[use,]),
  ensembleVerifObs(srftData[use,]))

k <- ensembleSize(srftData)

hist(srftVerifRank, breaks = 0:(k+1),
     prob = TRUE, xaxt = "n", xlab = "",
     main = "Verification Rank Histogram")
axis(1, at = seq(.5, to = k+.5, by = 1),
     labels = 1:(k+1))
abline(h=1/(ensembleSize(srftData)+1), lty=2)
```

The resulting rank histogram composites ranks spatially and is shown in Figure 5. The U shape indicates a lack of calibration, in that the ensemble forecast is underdispersed.

The PIT is the value that the predictive cumulative distribution function attains at the observation, and is a continuous analog of the verification rank. The function `pit` computes it. The PIT histogram allows for the visual assessment of calibration, and is interpreted in the same way as the verification rank histogram. We illustrate this on BMA forecasts of surface temperature obtained for the entire `srft` data set using a 25 day training period (forecasts begin on January 30, 2004 and end on February 28, 2004):

```
srftFitALL <- ensembleBMA( srftData,
                           trainingDays = 25)

srftPIT <- pit( srftFitALL, srftData)

hist(srftPIT, breaks = (0:(k+1))/(k+1),
     xlab="", xaxt="n", prob=TRUE,
     main = "Probability Integral Transform")
axis(1, at = seq(0, to = 1, by = .2),
     labels = (0:5)/5)
abline(h = 1, lty = 2)
```

The resulting PIT histogram is shown in Figure 5. It shows signs of negative bias, which is not surpising because it is based on only about a month of verifying data. We generally recommend computing the PIT histogram for longer periods, ideally at least a year, to avoid its being dominated by short-term and seasonal effects.

5 SUMMARY

We have described an R package called `ensembleBMA` for probabilistic weather forecasting, providing functionality to fit forecasting models to training data. In `ensembleBMA`, parametric mixture models, in which the components correspond to the members of an ensemble, are fit to a training set of ensemble forcasts, in a form that depends on the weather parameter. These models are then used to postprocess ensemble forecasts at station or grid locations. Supplementing model fitting and forecasting, the package also provides functionality for verification, allowing the quality of the forecasts produced to be assessed.

Acknowledgements. Supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745, by the National Science Foundation under grants ATM-0724721 and DMS-0706745, and by the Joint Ensemble Forecasting System (JEFS) under subcontract S06-47225 from the University Corporation for Atmospheric Research (UCAR). We are indebted to Cliff Mass and Jeff Baars of the University of Washington Department of Atmospheric Sciences for sharing insights, expertise and data, and thank Michael Scheuerer for comments.

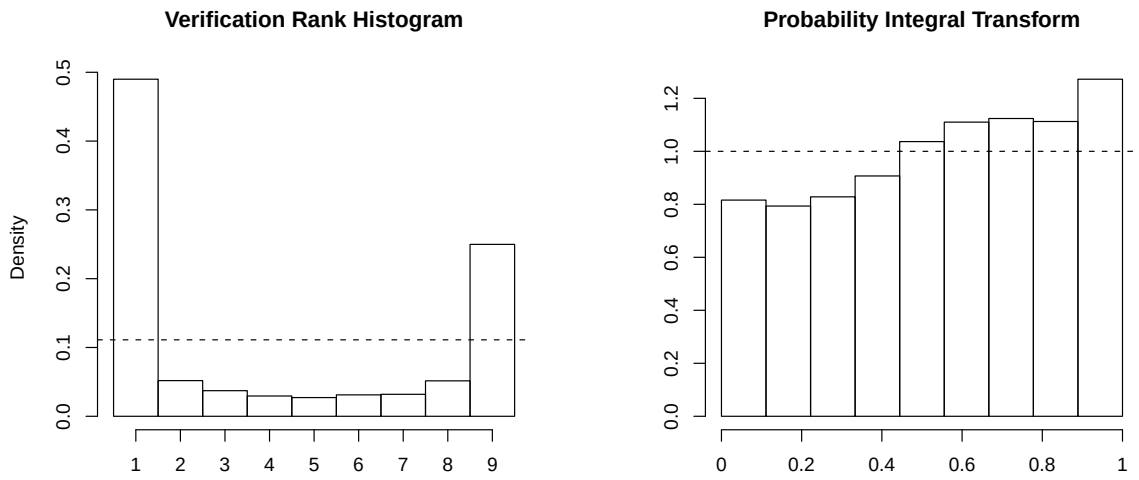

Figure 5: Verification rank histogram for ensemble forecasts, and PIT histogram for BMA forecast PDFs for surface temperature over the Pacific Northwest in the `srft` dataset valid from January 30, 2004 to February 28, 2004. More uniform histograms correspond to better calibration.

BIBLIOGRAPHY

Berrocal, V. J., A. E. Raftery, and T. Gneiting, 2007. Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Mon. Wea. Rev.*, 135:1386–1402.

Brownrigg, R., and T. P. Minka, 2003. maps: Draw geographical maps. (R package, revised 2010, original S code by R. A. Becker and A. R. Wilks).

Eckel, F. A., and C. F. Mass, 2005. Effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, 20:328–350.

Fraley, C., A. E. Raftery, T. Gneiting, and J. M. Sloughter, 2006. `ensembleBMA`: Probabilistic forecasting using ensembles and Bayesian Model Averaging. (R package, revised 2010).

Fraley, C., A. E. Raftery, T. Gneiting, and J. M. Sloughter, 2007. ENSEMBLEBMA: An R package for probabilistic forecasting using ensembles and Bayesian model averaging. Technical Report 516, University of Washington, Department of Statistics. (revised 2011).

Fraley, C., A. E. Raftery, and T. Gneiting, 2010. Calibrating multi-model forecasting ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.*, 138: 190–202.

Furrer, R., D. Nychka, and S. Sain, 2001. fields: Tools for spatial data. (R package, revised 2009).

Gel, Y. R., 2007. Comparative analysis of the local observation-based (LOB) method and the nonparametric regression-based method for gridded bias correction in mesoscale weather forecasting. *Wea. Forecasting*, 22:1243–1256.

Gneiting, T., and A. E. Raftery, 2005. Weather forecasting with ensemble methods. *Science*, 310: 248–249.

Gneiting, T., and A. E. Raftery, 2007. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, 102:359–378.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007. Probabilistic forecasts, calibration, and sharpness. *J. Roy. Stat. Soc., Ser. B*, 69:243–268.

Grimit, E. P., and C. F. Mass, 2002. Initial results for a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting*, 17:192–205.

Hamill,T. M.. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, 129:550–560.

Hornik, K., 1999. chron: Chronological objects which can handle dates and times. (R package, revised 2009, S original by D. James).

Jolliffe, I. T. and D. B. Stephenson, editors, 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, 2003.

Kleiber, W., A. E. Raftery, J. Baars, T. Gneiting, C. Mass, and E. P. Grimit, 2011. Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Mon. Wea. Rev.*, (in press).

Mass, C. F., J. Baars, G. Wedam, E. P. Grimit, and R. Steed, 2008. Removal of systematic model bias on a model grid. *Wea. Forecasting*, 23:438–459.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, 133:1155–1174.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, 135:3209–3220.

Sloughter, J. M., T. Gneiting, and A. E. Raftery, 2010. Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Amer. Stat. Assoc.*, 105:25–35.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007. Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Mon. Wea. Rev.*, 135:1364–1385.