Using Bayesian Model Averaging to Calibrate Forecast Ensembles

Adrian E. Raftery

Department of Statistics, University of Washington, Seattle

Joint work with Tilmann Gneiting (Washington), Fadoua Balabdaoui (Gottingen), Michael Polakowski (Oregon State) and Patrick Tewson (UW-APL)

Thanks to Cliff Mass, Eric Grimit, Jeff Baars, David Ovens, Mark Albright (UW-Atmos)

This work was supported by the U.S. DOD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745.

www.stat.washington.edu/raftery

www.stat.washington.edu/MURI

Outline

- Probabilistic forecasting using ensembles: They often show a spread-skill relationship but are uncalibrated

Outline

- Probabilistic forecasting using ensembles: They often show a spread-skill relationship but are uncalibrated
- Bayesian model averaging: A method based on statistical principles for producing probabilitistic forecasts from ensembles.
  It is *calibrated*, *sharp* and *honors the spread-skill relationship*

Outline

- Probabilistic forecasting using ensembles: They often show a spread-skill relationship but are uncalibrated
- Bayesian model averaging: A method based on statistical principles for producing probabilitistic forecasts from ensembles.
  It is *calibrated*, *sharp* and *honors the spread-skill relationship*
- Results for mesoscale forecasting in the Pacific Northwest

Some Definitions

- Probabilistic Forecast: A probability distribution of a future weather quantity or event

Some Definitions

- Probabilistic Forecast: A probability distribution of a future weather quantity or event
- Calibrated: Intervals or events that we declare to have probability P happen a proportion P of the time

Some Definitions

- Probabilistic Forecast: A probability distribution of a future weather quantity or event
- Calibrated: Intervals or events that we declare to have probability P happen a proportion P of the time
- Sharp: Prediction intervals are narrower on average than those obtained from climatology (i.e. the long run marginal distribution); the narrower the better

Some Definitions

- Probabilistic Forecast: A probability distribution of a future weather quantity or event
- Calibrated: Intervals or events that we declare to have probability P happen a proportion P of the time
- Sharp: Prediction intervals are narrower on average than those obtained from climatology (i.e. the long run marginal distribution); the narrower the better
- Goal: Maximize sharpness subject to calibration (Gneiting et al 2003)

Mesoscale Ensemble Forecasting

- The UW Mesoscale ensemble:

Mesoscale Ensemble Forecasting

- The UW Mesoscale ensemble:
  - A multianalysis ensemble started in January 2000 with 5 members (Phase I); now has 8+ members.

Mesoscale Ensemble Forecasting

- The UW Mesoscale ensemble:
  - A multianalysis ensemble started in January 2000 with 5 members (Phase I); now has 8+ members.
  - Each got by running MM5 with a different initialization, from a different global model and weather center

Mesoscale Ensemble Forecasting

- The UW Mesoscale ensemble:
  - A multianalysis ensemble started in January 2000 with 5 members (Phase I); now has 8+ members.
  - Each got by running MM5 with a different initialization, from a different global model and weather center
- Shows a clear spread-skill relationship, i.e. a correlation between the ensemble spread and the absolute error.

Mesoscale Ensemble Forecasting

- The UW Mesoscale ensemble:
  - A multianalysis ensemble started in January 2000 with 5 members (Phase I); now has 8+ members.
  - Each got by running MM5 with a different initialization, from a different global model and weather center
- Shows a clear spread-skill relationship, i.e. a correlation between the ensemble spread and the absolute error.
  - This spread-error correlation varies, but can reach up to 64% (Grimit & Mass 2002, Grimit 2004)

Lack of Calibration

- BUT it's not calibrated:

Lack of Calibration

- BUT it's not calibrated:
  - The ensemble range should have contained the truth about 67% of the time for the 5-member ensemble of Phase I

Lack of Calibration

- BUT it’s not calibrated:
  - The ensemble range should have contained the truth about 67% of the time for the 5-member ensemble of Phase I
  - but it did so only 29% of the time for 2m (surface) temperature

Lack of Calibration

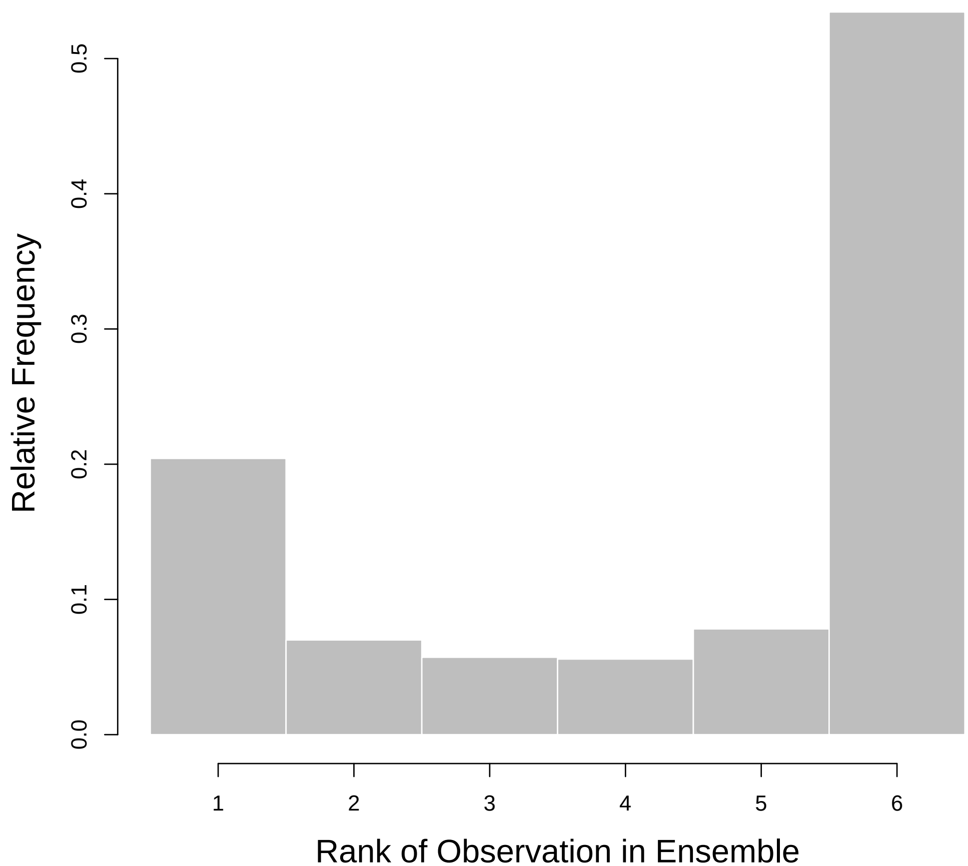
- BUT it's not calibrated:
  - The ensemble range should have contained the truth about 67% of the time for the 5-member ensemble of Phase I
  - but it did so only 29% of the time for 2m (surface) temperature
- Similar behavior observed with other ensembles, synoptic as well as mesoscale, particularly for surface parameters

Lack of Calibration

Verification rank histogram for surface temperature (should be uniform over the numbers 1,2,. . .,6):

Lack of Calibration

Verification rank histogram for surface temperature (should be uniform over the numbers 1,2,. . .,6):



Bayesian Model Averaging

- Standard statistical method for combining inferences (including predictions) from different models

Bayesian Model Averaging

- Standard statistical method for combining inferences (including predictions) from different models
- The overall (BMA) forecast probability distribution (PDF or CDF) is a weighted average of the forecast distributions from each model separately.

Bayesian Model Averaging

- Standard statistical method for combining inferences (including predictions) from different models
- The overall (BMA) forecast probability distribution (PDF or CDF) is a weighted average of the forecast distributions from each model separately.
- The weights are the estimated probabilities of the models, and reflect the models' predictive performance

Bayesian Model Averaging

- Standard statistical method for combining inferences (including predictions) from different models
- The overall (BMA) forecast probability distribution (PDF or CDF) is a weighted average of the forecast distributions from each model separately.
- The weights are the estimated probabilities of the models, and reflect the models' predictive performance
- The BMA point or deterministic forecast is just a weighted average of the forecasts in the ensemble.

Bayesian Model Averaging

- Standard statistical method for combining inferences (including predictions) from different models
- The overall (BMA) forecast probability distribution (PDF or CDF) is a weighted average of the forecast distributions from each model separately.
- The weights are the estimated probabilities of the models, and reflect the models' predictive performance
- The BMA point or deterministic forecast is just a weighted average of the forecasts in the ensemble.
- The BMA probability distribution can be represented as an *equally weighted* ensemble of any desired size, by simulating from the forecast distribution.

BMA for Mesoscale Forecasting at UW

- The predictive PDF is a mixture of five PDFs centered on the forecasts after bias correction.

BMA for Mesoscale Forecasting at UW

- The predictive PDF is a mixture of five PDFs centered on the forecasts after bias correction.
- Let y be the observed value.
  Let $\tilde{y}_k$ be the kth forecast.
  Then we have:

$$\begin{aligned} p(y|\tilde{y}_1, \ldots, \tilde{y}_5) &= w_1 N(a_1 + b_1 \tilde{y}_1, \sigma^2) + \ldots \\ &\quad + w_5 N(a_5 + b_5 \tilde{y}_5, \sigma^2), \end{aligned}$$

where $w_k \geq 0$, $\sum_{k=1}^{5} w_k = 1$.

BMA for Mesoscale Forecasting at UW

- The predictive PDF is a mixture of five PDFs centered on the forecasts after bias correction.
- Let y be the observed value.
  Let $\tilde{y}_k$ be the kth forecast.
  Then we have:

$$\begin{aligned} p(y|\tilde{y}_1, \ldots, \tilde{y}_5) &= w_1 N(a_1 + b_1\tilde{y}_1, \sigma^2) + \ldots \\ &\quad + w_5 N(a_5 + b_5\tilde{y}_5, \sigma^2), \end{aligned}$$

  where $w_k \geq 0$, $\sum_{k=1}^{5} w_k = 1$.
- The model is estimated from a training set of recent data by maximum likelihood using the EM algorithm. The estimate of σ^2 can be modified to minimize CRPS. Good results with a 25-day training period.

Example

48-Hour Forecast of Surface Temperature at Packwood, Wash. on June 12, 2000 at 00Z

Example

48-Hour Forecast of Surface Temperature at Packwood, Wash. on June 12, 2000 at 00Z

UW-MM5 Ensemble:

Example

48-Hour Forecast of Surface Temperature at Packwood, Wash. on June 12, 2000 at 00Z

UW-MM5 Ensemble:

Initialization	AVN	ETA	NGM	NOGAPS	GEM

Example

48-Hour Forecast of Surface Temperature at Packwood, Wash. on June 12, 2000 at 00Z

UW-MM5 Ensemble:

Initialization	AVN	ETA	NGM	NOGAPS	GEM
Source	(NCEP)	(NCEP)	(NCEP)	(FNMOC)	(MSC)

Example

48-Hour Forecast of Surface Temperature at Packwood, Wash. on June 12, 2000 at 00Z

UW-MM5 Ensemble:

Initialization	AVN	ETA	NGM	NOGAPS	GEM
Source	(NCEP)	(NCEP)	(NCEP)	(FNMOC)	(MSC)
Forecast	11	17	18	11	17

Example

48-Hour Forecast of Surface Temperature at Packwood, Wash. on June 12, 2000 at 00Z

UW-MM5 Ensemble:

Initialization	AVN	ETA	NGM	NOGAPS	GEM
Source	(NCEP)	(NCEP)	(NCEP)	(FNMOC)	(MSC)
Forecast	11	17	18	11	17

Observation: 19

Example (ctd)

- The observation was outside the ensemble range

Example (ctd)

- The observation was outside the ensemble range
- There was disagreement: 3 of the forecasts were around 17-18, and two were 11

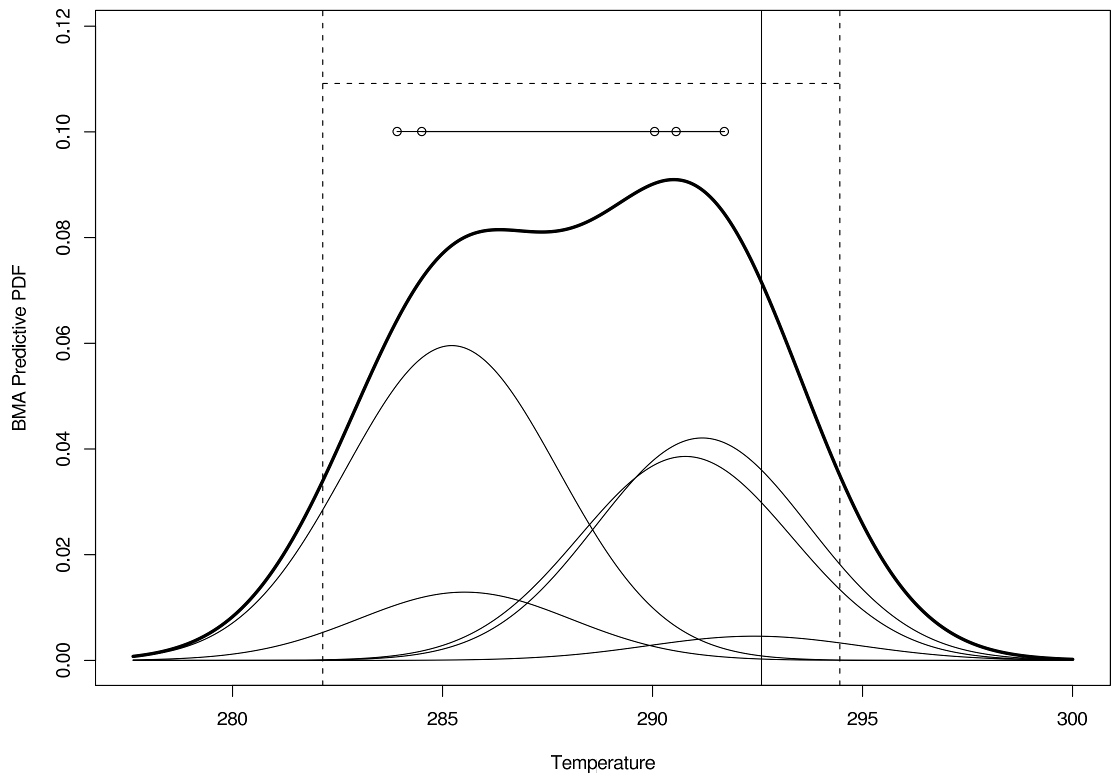
Example (ctd)

- The observation was outside the ensemble range
- There was disagreement: 3 of the forecasts were around 17-18, and two were 11

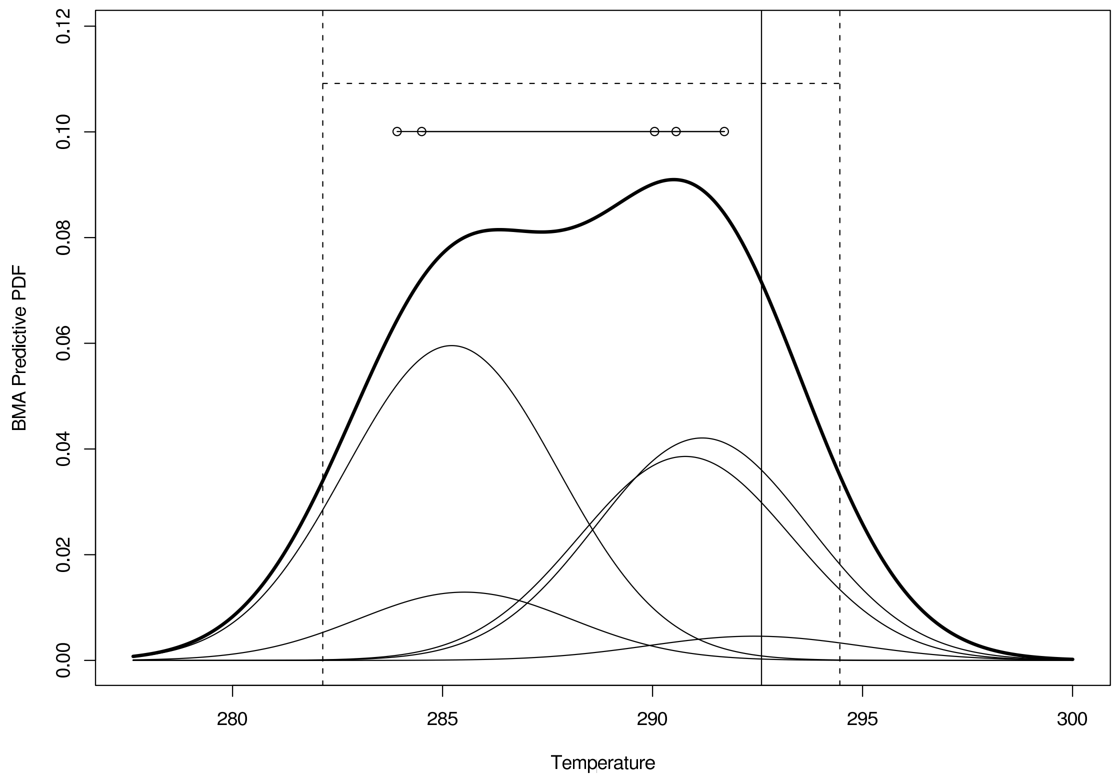
BMA Posterior Probabilities (%)

AVN (NCEP)	ETA (NCEP)	NGM (NCEP)	GEM (MSC)	NOGAPS (FNMOC)
38	27	3	24	8

BMA Forecast PDF



BMA Forecast PDF



The BMA forecast PDF is a weighted sum of 5 normals

BMA Forecast PDF

- The PDF has two “humps”, with one hump centered around the two lower forecasts, and the other hump centered around the three higher forecasts.

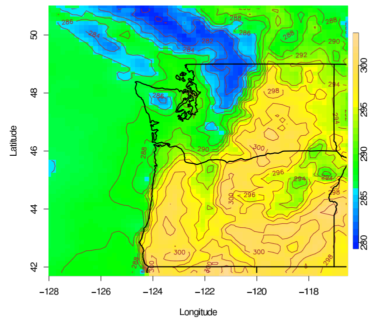
BMA Forecast PDF

- The PDF has two “humps”, with one hump centered around the two lower forecasts, and the other hump centered around the three higher forecasts.
- This reflects the disagreement among the forecasts

BMA Forecast PDF

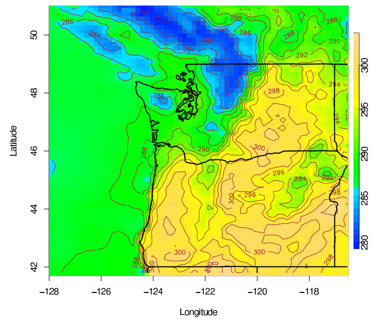
- The PDF has two “humps”, with one hump centered around the two lower forecasts, and the other hump centered around the three higher forecasts.
- This reflects the disagreement among the forecasts
- The observation falls in the 90% BMA forecast interval, although it is outside the ensemble range

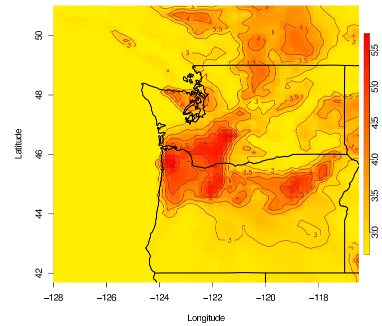
BMA Forecast and 90% Intervals



Deterministic Forecast

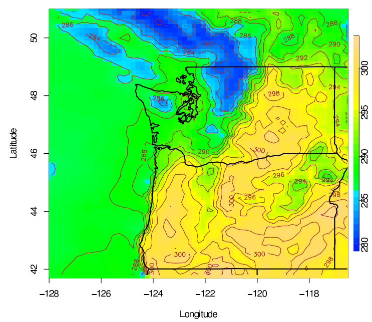
BMA Forecast and 90% Intervals

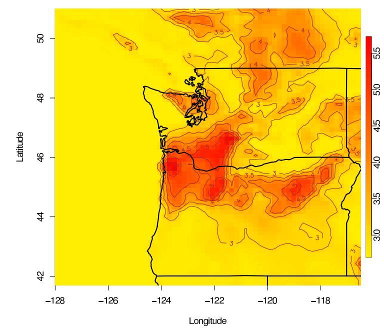




Deterministic Forecast Margin of Error

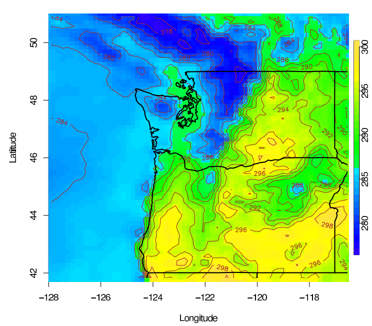
BMA Forecast and 90% Intervals





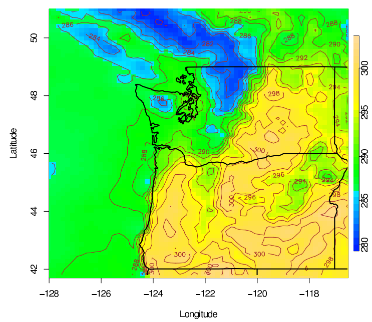
Deterministic Forecast

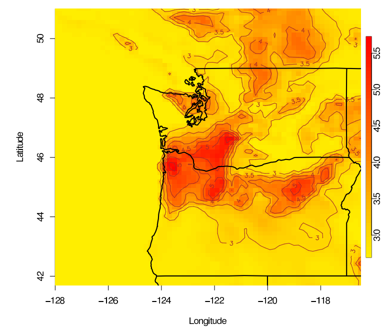
Margin of Error



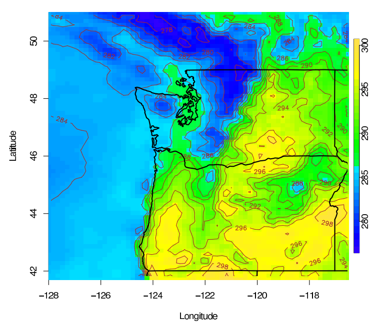
Lower bound

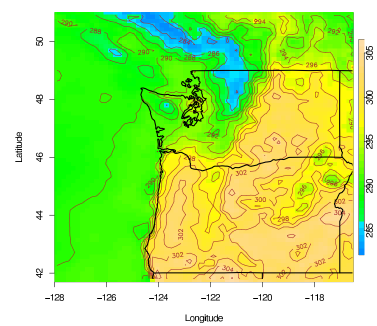
BMA Forecast and 90% Intervals





Deterministic Forecast Margin of Error





Lower bound Upper bound

Calibration of BMA Forecast Density

- We use the probability integral transform (PIT) histogram

Calibration of BMA Forecast Density

- We use the probability integral transform (PIT) histogram
- Continuous analogue of the verification rank histogram

Calibration of BMA Forecast Density

- We use the probability integral transform (PIT) histogram
- Continuous analogue of the verification rank histogram
- Results for surface temperature:

Calibration of BMA Forecast Density

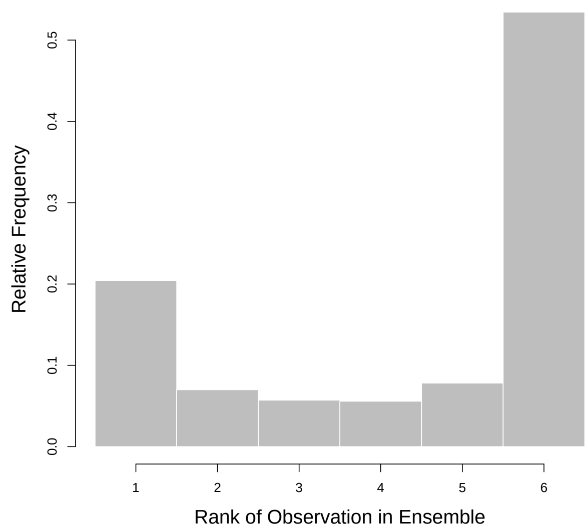
- We use the probability integral transform (PIT) histogram
- Continuous analogue of the verification rank histogram
- Results for surface temperature:

Ensemble VRH

Calibration of BMA Forecast Density

- We use the probability integral transform (PIT) histogram
- Continuous analogue of the verification rank histogram
- Results for surface temperature:

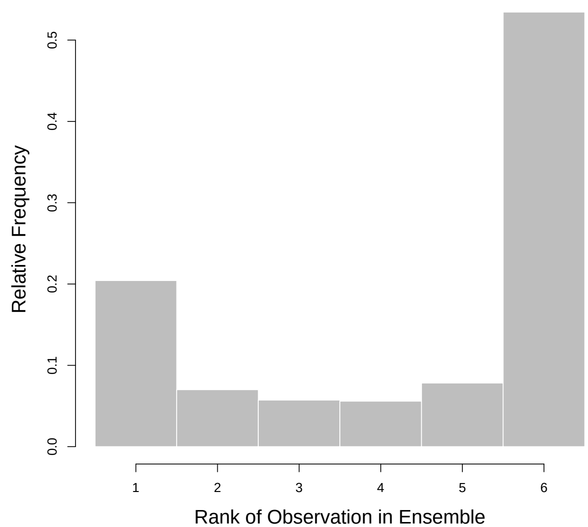
Ensemble VRH



Calibration of BMA Forecast Density

- We use the probability integral transform (PIT) histogram
- Continuous analogue of the verification rank histogram
- Results for surface temperature:

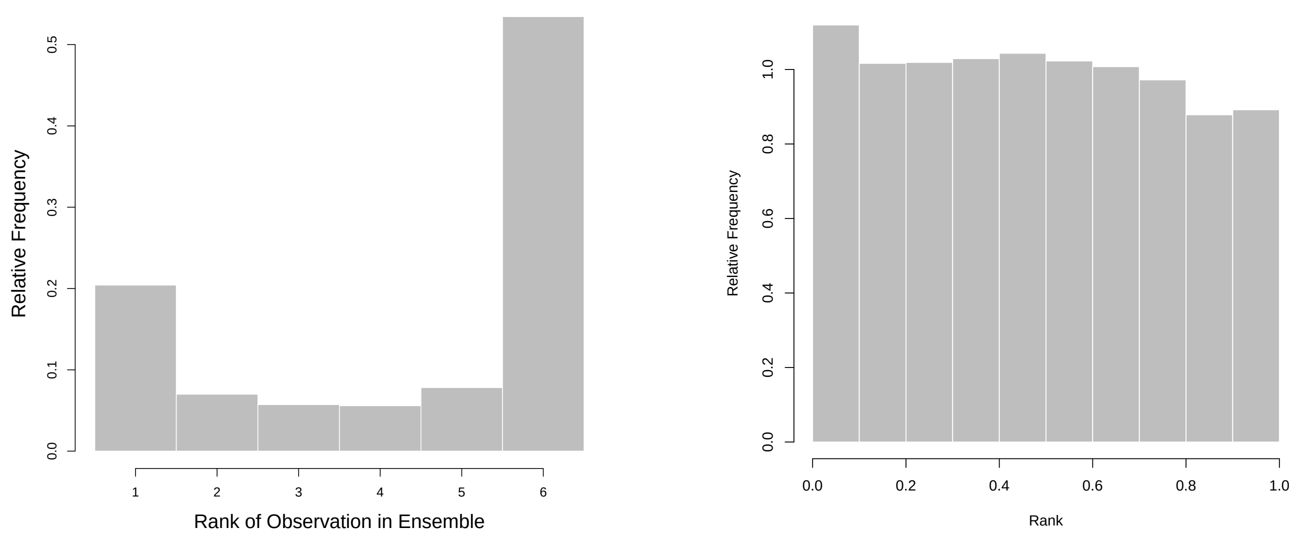
Ensemble VRH BMA PIT Histogram



Calibration of BMA Forecast Density

- We use the probability integral transform (PIT) histogram
- Continuous analogue of the verification rank histogram
- Results for surface temperature:

Ensemble VRH — BMA PIT Histogram



Verification Results: Calibration

(2m temperature)

$\frac{\text{Coverage of the 67\% prediction intervals}}{\text{(should be about 67\%):}}$

Verification Results: Calibration

(2m temperature)

Coverage of the 67% prediction intervals

(should be about 67%):

Sample climatology

Verification Results: Calibration

(2m temperature)
Coverage of the 67% prediction intervals
(should be about 67%):

Sample climatology	67%

Verification Results: Calibration

(2m temperature)
Coverage of the 67% prediction intervals
(should be about 67%):

Sample climatology	67%
Ensemble range	

Verification Results: Calibration

(2m temperature)
Coverage of the 67% prediction intervals

(should be about 67%):

Sample climatology	67%
Ensemble range	29%

Verification Results: Calibration

(2m temperature)
Coverage of the 67% prediction intervals
(should be about 67%):

Sample climatology	67%
Ensemble range	29%
BMA	

Verification Results: Calibration

(2m temperature)
Coverage of the 67% prediction intervals
(should be about 67%):

Sample climatology	67%
Ensemble range	29%
BMA	67%

Verification Results: Sharpness

Average width of the 67% prediction intervals ($^\circ$C) (smaller is better):

Verification Results: Sharpness

Average width of the 67% prediction intervals ($^\circ$C) (smaller is better):

Sample climatology

Verification Results: Sharpness

Average width of the 67% prediction intervals ($^\circ$C) (smaller is better):

Sample climatology	17.2

Verification Results: Sharpness

<u>Average width of the 67% prediction intervals ($^\circ$C)</u> (smaller is better):

Sample climatology	17.2
Ensemble range	

Verification Results: Sharpness

Average width of the 67% prediction intervals ($^\circ$C) (smaller is better):

Sample climatology	17.2
Ensemble range	2.5

Verification Results: Sharpness

Average width of the 67% prediction intervals ($^\circ$C) (smaller is better):

Sample climatology	17.2
Ensemble range	2.5
BMA	

Verification Results: Sharpness

Average width of the 67% prediction intervals ($^\circ$C) (smaller is better):

Sample climatology	17.2
Ensemble range	2.5
BMA	5.3

Verification Results: MAEs

MAEs of the deterministic forecasts ($^\circ$C)

(smaller is better):

Verification Results: MAEs

MAEs of the deterministic forecasts ($^\circ$C)

(smaller is better):

Sample climatology

Verification Results: MAEs

MAEs of the deterministic forecasts (°C) (smaller is better):

Sample climatology	7.7

Verification Results: MAEs

MAEs of the deterministic forecasts ($^\circ$C) (smaller is better):

Sample climatology	7.7
Best MM5 forecast	

Verification Results: MAEs

MAEs of the deterministic forecasts ($^\circ$C) (smaller is better):

Sample climatology	7.7
Best MM5 forecast	2.5

Verification Results: MAEs

MAEs of the deterministic forecasts ($^\circ$C) (smaller is better):

Sample climatology	7.7
Best MM5 forecast	2.5
Ensemble mean	

Verification Results: MAEs

MAEs of the deterministic forecasts ($^\circ$C)
(smaller is better):

Sample climatology	7.7
Best MM5 forecast	2.5
Ensemble mean	2.5

Verification Results: MAEs

MAEs of the deterministic forecasts (°C) (smaller is better):

Sample climatology	7.7
Best MM5 forecast	2.5
Ensemble mean	2.5
BMA	

Verification Results: MAEs

MAEs of the deterministic forecasts ($^\circ$C) (smaller is better):

Sample climatology	7.7
Best MM5 forecast	2.5
Ensemble mean	2.5
BMA	2.3

BMA and Spread-Skill

- Is BMA capturing the spread-skill relationship?

BMA and Spread-Skill

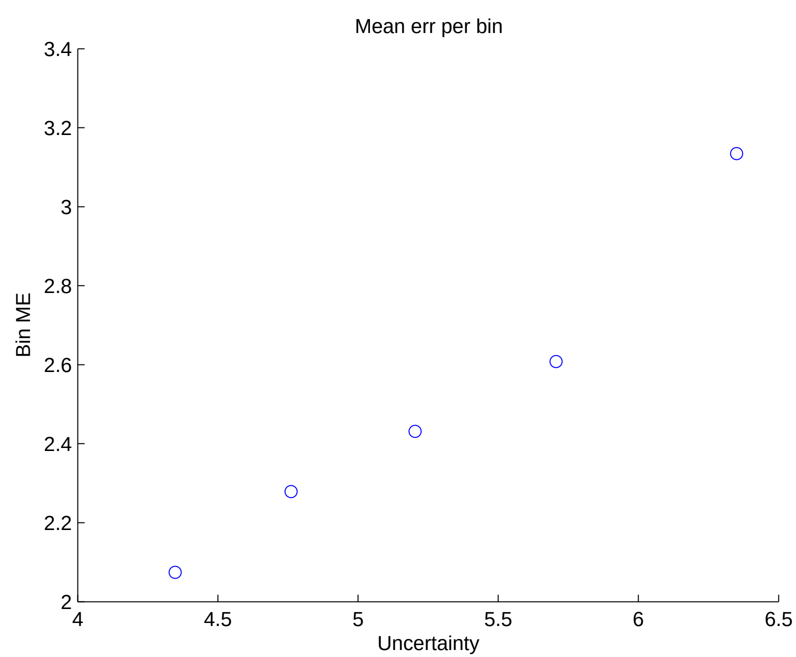
- Is BMA capturing the spread-skill relationship?
- I.e., is there a relationship between the width of the BMA interval and the absolute error?

BMA and Spread-Skill

- Is BMA capturing the spread-skill relationship?
- I.e., is there a relationship between the width of the BMA interval and the absolute error?
- We plot the forecast interval half-width (x-axis) against the mean absolute error (y-axis), for April–Nov 2004

BMA and Spread-Skill

- Is BMA capturing the spread-skill relationship?
- I.e., is there a relationship between the width of the BMA interval and the absolute error?
- We plot the forecast interval half-width (x-axis) against the mean absolute error (y-axis), for April–Nov 2004



Software: The EnsembleBMA Package

EnsembleBMA is a free software package written in the freely downloadable statistical language R.

Software: The EnsembleBMA Package

EnsembleBMA is a free software package written in the freely downloadable statistical language R.

Available at

`http://lib.stat.cmu.edu/R/CRAN/`

BMA at MSC and Elsewhere

- BMA is also being implemented by the Meteorological Service of Canada (MSC), German Weather Service (DWD) and Spanish Weather Service.

BMA at MSC and Elsewhere

- BMA is also being implemented by the Meteorological Service of Canada (MSC), German Weather Service (DWD) and Spanish Weather Service.
- At the North American Ensemble Workshop in Nov ’04, BMA was advocated as the ensemble postprocessing method of choice

BMA at MSC and Elsewhere

- BMA is also being implemented by the Meteorological Service of Canada (MSC), German Weather Service (DWD) and Spanish Weather Service.
- At the North American Ensemble Workshop in Nov ’04, BMA was advocated as the ensemble postprocessing method of choice
- Interest from NWS-Seattle, French, Japanese, Korean and New Zealand weather services

Current Research

- Extension to precip and wind (McLean Sloughter's talk)

Current Research

- Extension to precip and wind (McLean Sloughter's talk)
- Spatially coherent probabilistic forecasting: Bayesian dressing (Veronica Berrocal's talk)

Current Research

- Extension to precip and wind (McLean Sloughter's talk)
- Spatially coherent probabilistic forecasting: Bayesian dressing (Veronica Berrocal's talk)
- BMA implementation (Eric Grimit's talk):

Current Research

- Extension to precip and wind (McLean Sloughter's talk)
- Spatially coherent probabilistic forecasting: Bayesian dressing (Veronica Berrocal's talk)
- BMA implementation (Eric Grimit's talk):
  - Use observations as truth or an analysis?

Current Research

- Extension to precip and wind (McLean Sloughter's talk)
- Spatially coherent probabilistic forecasting: Bayesian dressing (Veronica Berrocal's talk)
- BMA implementation (Eric Grimit's talk):
  - Use observations as truth or an analysis?
  - How do BMA parameters vary in space?

Current Research

- Extension to precip and wind (McLean Sloughter's talk)
- Spatially coherent probabilistic forecasting: Bayesian dressing (Veronica Berrocal's talk)
- BMA implementation (Eric Grimit's talk):
  - Use observations as truth or an analysis?
  - How do BMA parameters vary in space?
- Displaying probabilistic forecasts: The UW Ensemble BMA web page (Patrick Tewson's talk)

Current Research

- Extension to precip and wind (McLean Sloughter's talk)
- Spatially coherent probabilistic forecasting: Bayesian dressing (Veronica Berrocal's talk)
- BMA implementation (Eric Grimit's talk):
  - Use observations as truth or an analysis?
  - How do BMA parameters vary in space?
- Displaying probabilistic forecasts: The UW Ensemble BMA web page (Patrick Tewson's talk)
- Better bias correction (Cliff Mass, Rick Steed)

Messages

- Forecast ensembles tend to show a spread-skill relationship, but still be underdispersed

Messages

- Forecast ensembles tend to show a spread-skill relationship, but still be underdispersed
- Bayesian model averaging is a statistical way of getting sharp calibrated probabilistic forecasts from an ensemble, that honor the spread-skill relationship

Messages

- Forecast ensembles tend to show a spread-skill relationship, but still be underdispersed
- Bayesian model averaging is a statistical way of getting sharp calibrated probabilistic forecasts from an ensemble, that honor the spread-skill relationship
- In experiments with temperature and pressure in the Pacific Northwest, BMA was calibrated, sharp, and gave good deterministic forecasts

Messages

- Forecast ensembles tend to show a spread-skill relationship, but still be underdispersed
- Bayesian model averaging is a statistical way of getting sharp calibrated probabilistic forecasts from an ensemble, that honor the spread-skill relationship
- In experiments with temperature and pressure in the Pacific Northwest, BMA was calibrated, sharp, and gave good deterministic forecasts
- Free R package: EnsembleBMA at

Messages

- Forecast ensembles tend to show a spread-skill relationship, but still be underdispersed
- Bayesian model averaging is a statistical way of getting sharp calibrated probabilistic forecasts from an ensemble, that honor the spread-skill relationship
- In experiments with temperature and pressure in the Pacific Northwest, BMA was calibrated, sharp, and gave good deterministic forecasts
- Free R package: EnsembleBMA at

`http://lib.stat.cmu.edu/R/CRAN/`

Messages

- Forecast ensembles tend to show a spread-skill relationship, but still be underdispersed
- Bayesian model averaging is a statistical way of getting sharp calibrated probabilistic forecasts from an ensemble, that honor the spread-skill relationship
- In experiments with temperature and pressure in the Pacific Northwest, BMA was calibrated, sharp, and gave good deterministic forecasts
- Free R package: EnsembleBMA at

  `http://lib.stat.cmu.edu/R/CRAN/`

- www.stat.washington.edu/raftery
  www.stat.washington.edu/MURI