Assessing predictive distributions: a diagnostic approach

Tilmann Gneiting and Adrian E. Raftery

University of Washington
Department of Statistics

joint work with Fadoua Balabdaoui, Yulia Gel and Anton Westveld

supported by **DoD Multidisciplinary University Research Initiative (MURI)**

presented 31 July 2002 at **NCAR Verification Workshop**

Outline

University of Washington MURI project

University of Washington MM5 mesoscale short-range ensemble forecasting system

From ensemble forecasts to predictive distributions

Assessing predictive distributions:
a diagnostic approach

- assessing calibration
- assessing sharpness
- scores/summary measures

Discussion

University of Washington MURI project

DoD
Multidisciplinary **U**niversity **R**esearch **I**nitiative

goal: to develop ways of assessing and communicating uncertainty in mesoscale numerical weather forecasts

interdisciplinary: collaboration between the Departments of Atmospheric Sciences, Statistics, and Psychology, and the Applied Physics Laboratory

meteorology	**data and underlying science** Cliff Mass
statistics	**uncertainty assessment** Adrian Raftery Tilmann Gneiting
psychology	**communication of uncertainty** Earl ("Buz") Hunt Susan Joslyn
applied physics	**visualization systems** Robert Miyamoto David Jones Scott Sandgaithe

common challenge: **verification**

University of Washington MM5 mesoscale short-range ensemble forecasting system

Eric Grimit and Cliff Mass

www.atmos.washington.edu/~epgrimit/ensemble.cgi

Western North America and NE Pacific Ocean, 36-km domain

Washington, Oregon and S British Columbia, 12-km domain

phase I ensemble consists of **MM5 runs driven by five different global models:** AVN, CMC, ETA, NGM, and NOGAPS

supported by a consortium of local and federal agencies

From ensemble forecasts to predictive distributions

Predictive distributions

univariate, continuous predictand X
for example, windspeed at a certain location

but could be any user-relevant functional of the model output

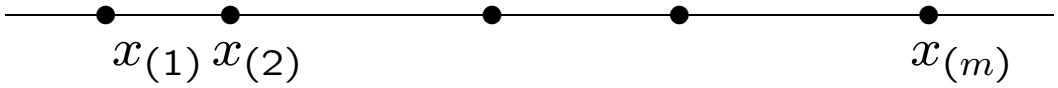
seek a **probabilistic forecast** in the form of a **predictive distribution function**

$$F(x) = P(X < x) \quad \text{for} \quad x \in \mathbb{R}$$

Natural approach based on an ensemble forecast system

ensemble values $x_1, \ldots, x_m$ (phase I: $m = 5$)

their order statistics $x_{(1)} < x_{(2)} < \cdots < x_{(m)}$ partition the real line into $m + 1$ bins



and we obtain a crude predictive distribution through the formula

$$F(x_{(i)}) = \frac{i}{m+1} \quad \text{for} \quad i = 1, \ldots, m$$

specified at m points only

Richardson (2001) and Palmer (2002): lessens the economic value of an ensemble forecast system

Full predictive distributions through statistical postprocessing

Hamill and Colucci (1998), Eckel and Walters (1998):

- stratification by ensemble spread
- calibration
- linear interpolation and climatological extrapolation of the natural approach

Wilson, Burrows and Lanzinger (1999):

- fit parametric predictive distributions to the ensemble output

Raftery, Balabdouai and Gneiting (200x):

- **BMA: B**ayesian **M**odel **A**veraging

Gneiting, Westveld and Raftery (200x):

EMOS: Ensemble **M**odel **O**utput **S**tatistics

multiple regression model

$$X = a + b_1 x_1 + \cdots + b_m x_m + e$$

where $e \sim \mathcal{N}(0, \sigma^2)$; basically

$$F = \mathcal{N}\left(a + b_1 x_1 + \cdots + b_m x_m,\ \sigma^2\right)$$

CHMOS: Conditionally **H**eteroscedastic **M**odel **O**utput **S**tatistics

modification in which

$$F = \mathcal{N}\left(a + b_1 x_1 + \cdots + b_m x_m,\ (c^2 S^2 + 1)\,\sigma^2\right)$$

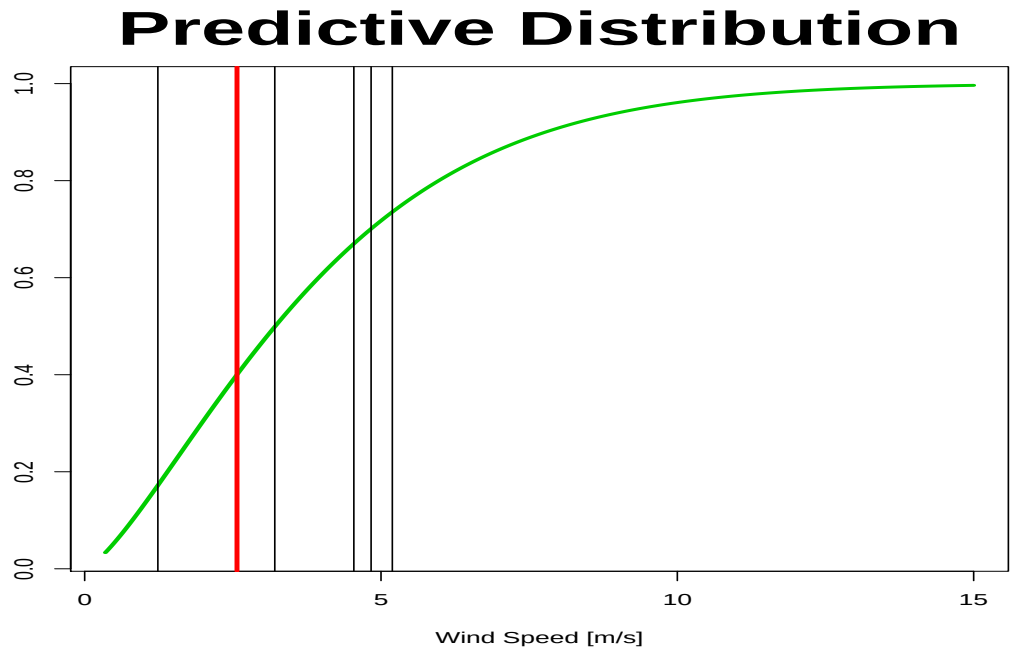
respects spread-error relationship

uncertainty decomposition: between-model (IC) and within-model (physics) uncertainty

Example

probabilistic wind speed forecasts based on the UW MM5 ensemble forecast system

phase I data, training and test set



48-hour predictive distribution for wind speed W of La Push, Olympic Peninsula, verified January 16, 2000, 4pm

various approaches of forming predictive distributions

need to **compare and assess**

a single probabilistic forecast cannot be verified: need to average

Assessing predictive distributions: a diagnostic approach

Requirements

calibration (reliability, statistical consistency): we want events with forecast probability p to verify with relative frequency p

sharpness (refinement): we want to minimize the spread of the predictive distributions

Key points

diagnostic approach as proposed and implemented by Murphy, Brown and Chen (1989) in the deterministic context

calibration diagram and **sharpness diagram**

summary measures or **scores don't suffice**

Assessing calibration

predictive distribution F_t for $t = 1, 2, \ldots$
verified value or **observation** x_t for $t = 1, 2, \ldots$

define the **probability rank**

$$p_t = F_t(x_t) \quad \text{for} \quad t = 1, 2, \ldots$$

predictive distributions are **probabilistically calibrated** relative to the observations if

$$\frac{1}{T} \#\{p_t < p : t = 1, \ldots, T\} \to p \text{ for all } p \in (0, 1)$$

calibration is a **joint property of predictive distributions and observations**

Hoeting diagram

plot of observed frequency of probability ranks,

$$\frac{1}{T}\,\#\{p_t < p : t = 1, \ldots, T\} \quad \text{versus} \quad p$$

should be a straight line

proposed by Hoeting (1994) and recently used by Moyeed and Papritz (2002)

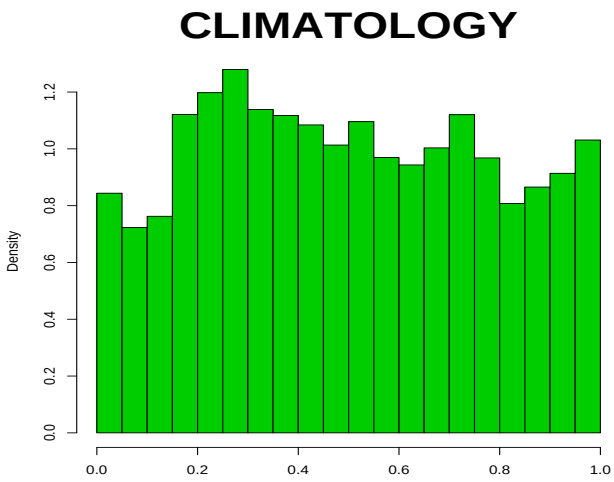
Calibration diagram

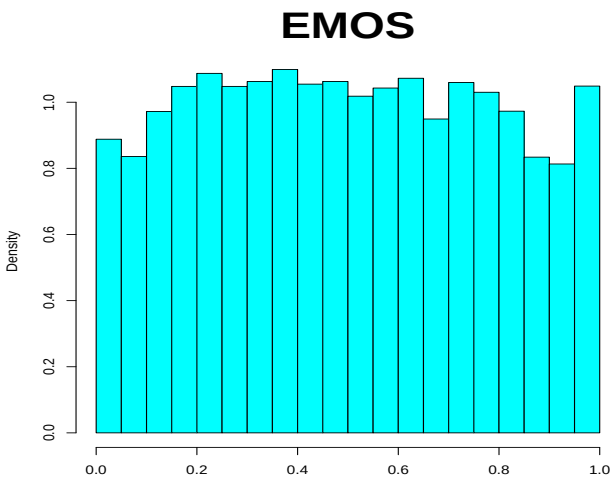
histogram of the probability ranks p_t

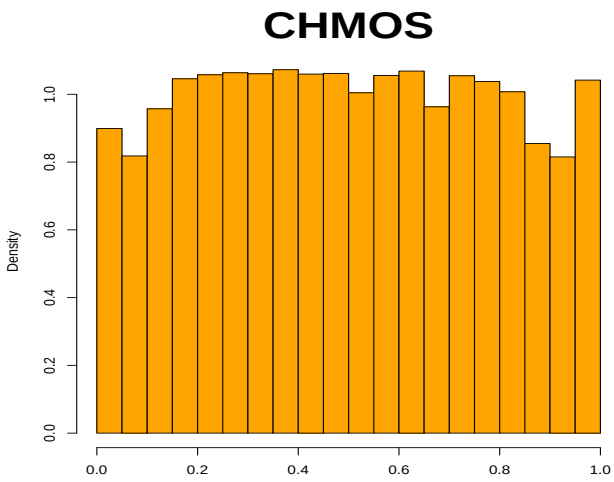
should be flat

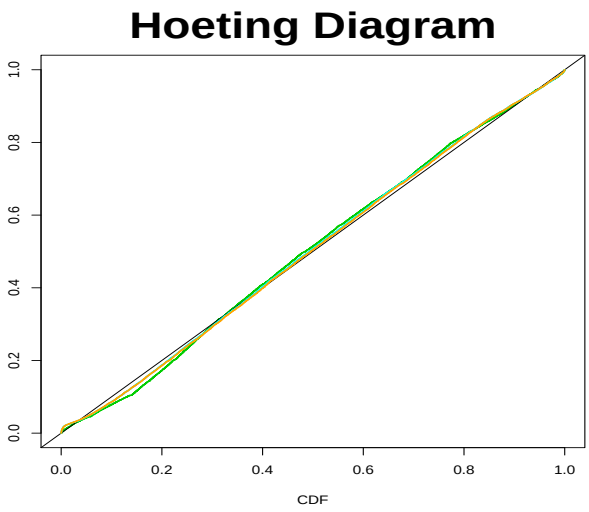
similar to the **verification rank histogram** or **Talagrand diagram** (Anderson 1996, Talagrand, Vautard and Strauss 1996, Hamill and Colucci 1997)

and to the **multicategory reliability diagram** (Hamill 1997)









suggestions:

- violations of calibration easier to detect in histogram form
- choosing 20 bins in the calibration diagram works well

Assessing sharpness

sharpness refers to the spread of the predictive distributions

should be as small as possible

a **property of the predictive distributions only**

challenge of probabilistic prediction is to **maximize sharpness while maintaining calibration**

Sharpness plot

key tool to assess sharpness

displays the spread of the predictive distributions: best understood by example

Probabilistic wind speed forecasts in the Pacific Northwest

predictive distribution $F_t(x) = P(X_t < x)$ for $t = 1, \ldots, T$

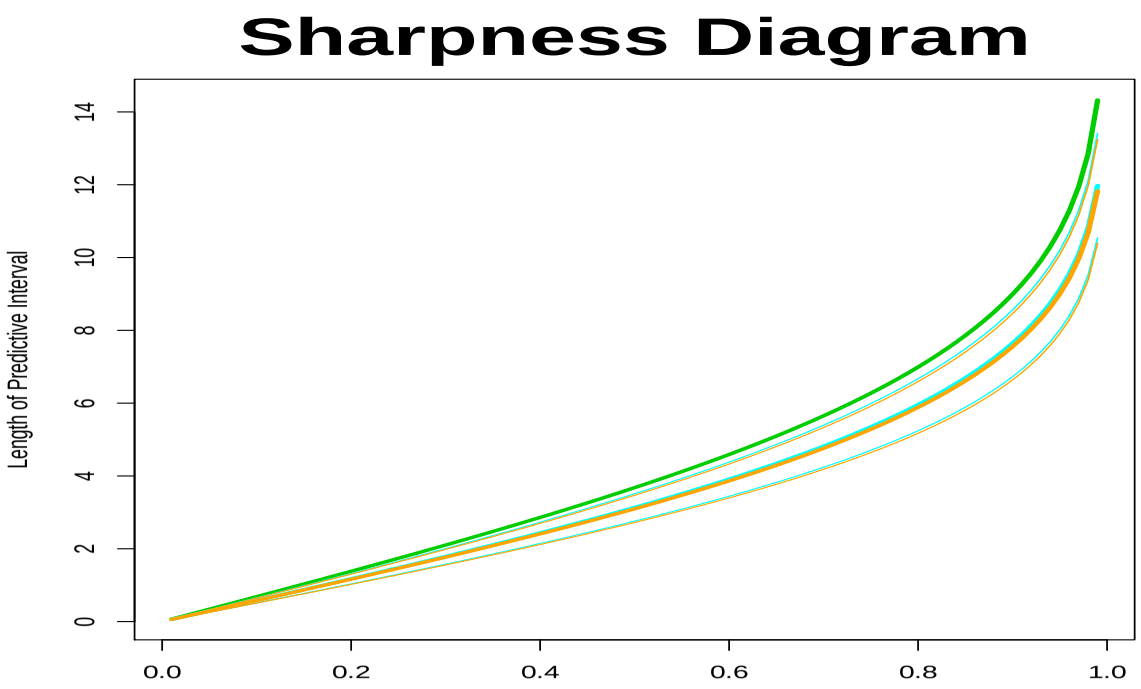
given $p \in (0, 1)$, we define the **predictive quantile** $q_{p,t}$ by

$$F_t(q_{t,p}) = p$$

given $\alpha \in (0, 1)$, the **length** of the **α-level central predictive interval** is

$$l_t(\alpha) = q_{t,\frac{1+\alpha}{2}} - q_{t,\frac{1-\alpha}{2}}$$

the **sharpness plot** displays $l_t(\alpha)$ versus α


Sharpness Diagram
Length of Predictive Interval
0
2
4
6
8
10
12
14
0.0
0.2
0.4
0.6
0.8
1.0

Scores/summary measures

predictive distribution $F_t(\cdot)$ with **predictive density** $f_t(\cdot)$, for $t = 1, \ldots, T$

verified value or **observation** x_t

evaluation in terms of a **score** or **summary measure** for $(F_t(\cdot), x_t)$ or $(f_t(\cdot), x_t)$

Traditional scores

spherical, log, and **quadratic score**

$$\begin{aligned} \mathsf{SphS}(f_t(\cdot), x_t) &= f_t(x) \Big/ \Big(\textstyle\int_{-\infty}^{\infty} f_t^2(u)\, \mathrm{d}u \Big)^{1/2} \\ \mathsf{LogS}(f_t(\cdot), x_t) &= \log f_t(x) \\ \mathsf{QS}(f_t(\cdot), x) &= 2\, f_t(x) - \textstyle\int_{-\infty}^{\infty} f_t^2(u)\, \mathrm{d}u \end{aligned}$$

continuous ranked probability score

$$\mathsf{CRPS}(F_t(\cdot), x_t) = \int_{-\infty}^{\infty} \left(F_t(u) - \mathbf{1}\{u > x_t\}\right)^2 \mathrm{d}u$$

I. J. Good class

$$\text{GoodS}_\beta(f_t(\cdot), x_t) = \frac{1}{\beta - 1}\left(\left(\frac{f_t(x_t)}{\left(\int_{-\infty}^{\infty} f_t^\beta(u)\,\mathrm{d}u\right)^{1/\beta}}\right)^{\beta-1} - 1\right)$$

$\beta = 2$ spherical score

$\beta \to 1$ log score

Properties of the scores

strictly proper, i.e., the forecaster maximizes her expected score if she states her true beliefs

but **not** combined measures of calibration and sharpness

supplement but **not a substitute** for calibration and sharpness diagram

key use in probabilistic forecast competitions

Discussion

- statistical postprocessing of ensemble forecasts yields full **predictive distributions**
- **verification** of predictive distributions is a challenging and largely unexplored endeavor
- **diagnostic approach** is essential: need to assess both **calibration** and **sharpness**
- key tools: **calibration diagram** and **sharpness diagram**

Addendum: probabilistic calibration and exceedance calibration

predictive distribution $F_t(\cdot)$ for $t = 1, 2, \ldots$
verified value or **observation** x_t for $t = 1, 2, \ldots$

predictive distributions are

probabilistically calibrated relative to the observations if

$$\frac{1}{T} \#\{p_t < p : t = 1, \ldots, T\} \to p \text{ for all } p \in (0, 1)$$

where $p_t = F_t(x_t)$ is the **probability rank**

exceedance calibrated relative to the observations if

$$\frac{1}{T} \#\{x_t < x : t = 1, \ldots, T\} \to \bar{F}(x) \text{ for all } x \in \mathbb{R}$$

where $\bar{F}(x) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} F_t(x)$

calibrated relative to the observations if they are both probabilistically calibrated and exceedance calibrated