Assessing Probabilistic Forecasts of Continuous Weather Variables

Tilmann Gneiting
University of Washington

Int'l Verification Methods Workshop
15 September 2004

joint work with Adrian E. Raftery, Fadoua Balabdaoui, Kristin Larson, Kenneth Westrick, Marc G. Genton and Eric Aldrich

University of Washington, 3TIER Environmental Forecast Group, Inc. and Texas A&M University

supported by DoD Multidisciplinary University Research Initiative (MURI), WTC and NSF

Probabilistic forecasts

Calibration and sharpness

Scoring rules

Case study:
Short-range forecasts of wind speed

Probabilistic forecasts

univariate, continuous or mixed discrete continuous **predictand** X

probabilistic forecast in the form of a predictive **cumulative distribution function (CDF)**

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}$$

or a predictive **probability density function (PDF)**

$$f(x), \quad x \in \mathbb{R}$$

examples include

raw **ensemble forecasts** of temperature, pressure, precipitation, wind speed, . . .

postprocessed ensemble forecasts (ensemble smoothing, BMA, EMOS)

statistical short-range forecasts of wind speed at wind energy sites

What is a good probabilistic forecast?

ECMWF Workshop on Predictability 1997:

> . . . the primary purpose of the ensemble prediction system is to provide and estimate the probability density function (pdf) of the atmospheric state. Such a pdf should possess two properties:
>
> 1. **statistical consistency** (or **reliability**)
>
> 2. **usefulness**, that is, the pdf should provide more accurate information about the predicted atmospheric state than a reference pdf based either on climatology or on a combination of deterministic (operational) forecasts and statistical data.

Calibration and sharpness

calibration:

statistical compatibility between the predictive distributions and the observations

joint property of the forecasts and the observations

sharpness:

refers to the spread of the predictive distributions

property of the forecasts only

goal of probabilistic forecasting is to **maximize sharpness subject to calibration**

Game-theoretic framework

two players, **nature** and **forecaster**

prequential scenario: times (cases, locations, ...) $t = 1, 2, \ldots$

nature chooses a distribution G_t

forecaster chooses a distribution F_t

the **observation** or **verification** x_t is a **random draw** from G_t

verification on the basis of the (F_t, x_t)

Example

at time t, nature chooses

$$G_t = \mathcal{N}(\mu_t, 1) \quad \text{where} \quad \mu_t \sim \mathcal{N}(0, 1)$$

perfect forecaster

$$F_t = G_t = \mathcal{N}(\mu_t, 1) \quad \text{for all } t$$

climatological forecaster

$$F_t = \mathcal{N}(0, 2) \quad \text{for all } t$$

Tom Hamill's forecaster

$$F_t = \begin{cases} \mathcal{N}\left(\mu_t - \frac{1}{2}, 1\right) & \text{with probability } \frac{1}{3} \\ \mathcal{N}\left(\mu_t + \frac{1}{2}, 1\right) & \text{with probability } \frac{1}{3} \\ \mathcal{N}\left(\mu_t, \left(\frac{13}{10}\right)^2\right) & \text{with probability } \frac{1}{3} \end{cases}$$

Notions of calibration

probabilistic calibration

$$\frac{1}{T}\sum_{t=1}^{T} G_t\left(F_t^{-1}(p)\right) \longrightarrow p \quad \text{for all } p \in (0,1)$$

exceedance calibration

$$\frac{1}{T}\sum_{t=1}^{T} G_t^{-1}\left(F_t(x)\right) \longrightarrow x \quad \text{for all } x$$

marginal calibration

$$\frac{1}{T}\sum_{t=1}^{T} \left(G_t(x) - F_t(x)\right) \longrightarrow 0 \quad \text{for all } x$$

perfect forecaster: PEM
climatological forecaster: P$\overline{\text{E}}$M
Hamill's forecaster: P$^*\overline{\text{E}}\,\overline{\text{M}}$

Verification tools

verification based on (F_t, x_t)

Assessing probabilistic calibration

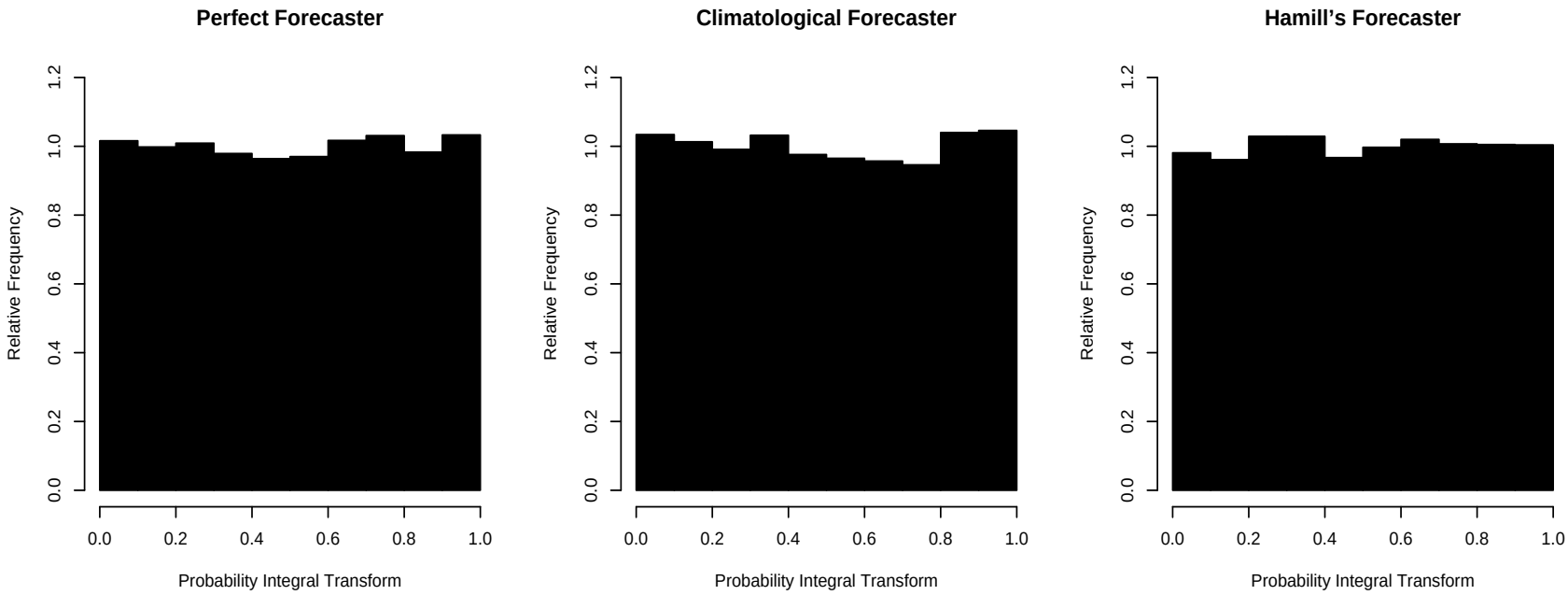
probability integral transform or **PIT** (Rosenblatt 1952; Dawid 1984)

$$p_t = F_t(x_t) \in [0, 1]$$

PIT histogram: histogram of the p_t

PIT histogram uniform $\iff$ prediction intervals **at all levels** have proper coverage

analogue of the **verification rank histogram** for ensemble forecasts


Perfect Forecaster
Relative Frequency
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.2
0.4
0.6
0.8
1.0
Probability Integral Transform
Climatological Forecaster
Relative Frequency
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.2
0.4
0.6
0.8
1.0
Probability Integral Transform
Hamill's Forecaster
Relative Frequency
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.2
0.4
0.6
0.8
1.0
Probability Integral Transform

Assessing marginal calibration

addresses compatibility between verifying climatology and forecast climatology

histogram of the verifications x_t

for each t, draw a random number y_t from the predictive distribution F_t

histogram of the y_t

marginal calibration table compares 5%, 50% and 95% percentiles of the histograms

	5%	50%	95%
Verifications	−2.37	0.01	2.31
Perfect forecaster	−2.28	0.00	2.30
Climatological forecaster	−2.34	0.02	2.37
Hamill's forecaster	−2.59	0.02	2.64

Assessing sharpness

average width of **90% central prediction interval**

	Ave Width
Perfect forecaster	3.29
Climatological forecaster	4.65
Hamill's forecaster	3.62

Scoring rules

a **scoring rule**

$$S(F, x)$$

assigns a numerical score to the forecast/observation pair (F, x)

negatively oriented: we consider scores to be penalties

the smaller the better: the forecaster aims to minimize the average score,

$$\frac{1}{T} \sum_{t=1}^{T} S(F_t, x_t)$$

diagnostic approach: scoring rules address both calibration and sharpness, yet form one facet of forecast verification only

Propriety

suppose that I provide probabilistic forecasts of a real-valued quantity X for your company

my best assessment: G
my actual forecast: F
verification: x
my penalty: $S(F, x)$

you expect me to quote $F = G$; however, will I do so?

only if the expected score is minimized if I quote $F = G$, that is if

$$\mathsf{E}_G\, S(G, X) \leq \mathsf{E}_G\, S(F, X)$$

for all F and G

a scoring rule with this property is called **proper**

all scoring rules discussed hereinafter are proper

Scoring rules for PDF forecasts

ignorance score (Good 1952; Roulston and Smith 2002)

$$\mathrm{IgnS}(f, x) = -\log f(x)$$

specifically,

$$\mathrm{IgnS}(\mathcal{N}(\mu, \sigma^2), x) = \frac{1}{2}\ln(2\pi\sigma^2) + \frac{(x-\mu)^2}{2\sigma^2}$$

quadratic score and **spherical score** (Good 1971)

$$\mathrm{QS}(f, x) = -f(x) + \frac{1}{2}\int_{-\infty}^{\infty} (f(y))^2 \, \mathrm{d}y$$

$$\mathrm{SphS}(f, x) = -f(x) \Big/ \left(\int_{-\infty}^{\infty} (f(y))^2 \, \mathrm{d}y\right)^{1/2}$$

Scoring rules for predictice CDFs

the **continuous ranked probability score** or **CRPS** has lately attracted attention

origins unclear (Matheson and Winkler 1976; Staël von Holstein 1977; Unger 1985)

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} \left(F(y) - \mathbf{1}(y \geq x)\right)^2 \, \mathrm{d}y$$

integral of the Brier scores for probability forecasts at all possible threshold values y

specifically,

$$\mathrm{CRPS}\Big(\mathcal{N}(\mu, \sigma^2), x\Big) = \sigma \left(\frac{x-\mu}{\sigma} \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right) + 2\,\varphi\left(\frac{x-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}} \right)$$

grows linearly in $|x - \mu|$, in contrast to the ignorance score

using results of Székely (2003)

$$\begin{aligned} \mathrm{CRPS}(F, x) &= \int_{-\infty}^{\infty} (F(y) - \mathbf{1}(y > x))^2 \, \mathrm{d}y \\ &= \mathrm{E}_F \, |X - x| - \frac{1}{2} \mathrm{E}_F \left| X - X' \right| \end{aligned}$$

where X and X' are independent random variables, both with distribution F

generalizes the absolute error to which it reduces if F is a deterministic (point) forecast

can be reported in the **same unit as the verifications**

provides a **direct way of comparing deterministic and probabilistic forecasts**

forms a special case of a novel and very general type of score, the **energy score** (Gneiting and Raftery 2004)

Scores for quantile and interval forecasts

consider **interval forecasts** in the form of the **central $(1-\alpha) \times 100\%$ prediction intervals**

equivalent to **quantile forecasts** at the **levels $\frac{\alpha}{2} \times 100\%$** and **$(1-\frac{\alpha}{2}) \times 100\%$**

$\alpha = 0.10$ corresponds to the 90% central prediction interval and quantile forecasts at the 5% and 95% level

scoring rule $S_\alpha(l, u; x)$ if the interval forecast is $[l, u]$ and the verification is x

interval score

$$S_\alpha(l,u;x) = \begin{cases} 2\alpha(u-l) + 4(l-x) & \text{if } x < l \\ 2\alpha(u-l) & \text{if } x \in [l,u] \\ 2\alpha(u-l) + 4(x-u) & \text{if } x > u \end{cases}$$

fixed penalty proportional to width of interval

additional penalty if the verification falls outside the prediction interval

Case study:
Short-range forecasts of wind speed

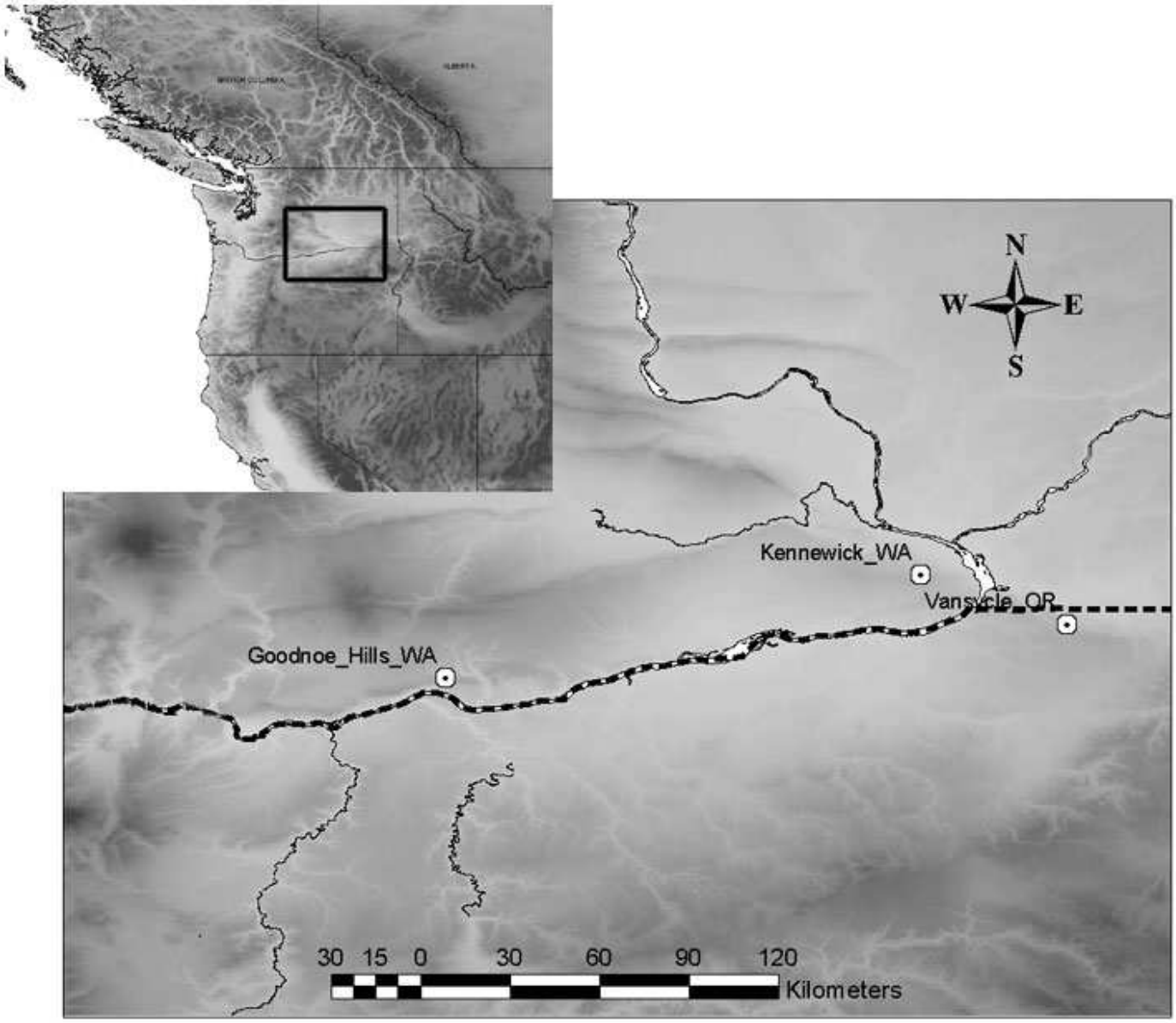
wind power: the world's fastest growing energy source; clean and renewable

Stateline wind energy center: $300 million wind project on the **Vansycle ridge** at the Oregon-Washington border

2-hour forecasts of hourly average **wind speed at the Vansycle ridge**

joint project with **3TIER Environmental Forecast Group, Inc.**

data collected by Oregon State University for the Bonneville Power Administration


N
W
E
S
Kennewick_WA
Vansycle_OR
Goodnoe_Hills_WA
30 15 0 30 60 90 120
Kilometers

Forecast techniques

persistence forecast as reference standard:

$$\widehat{V}_{t+2} = V_t$$

classical approach (Brown, Katz and Murphy 1984): **autoregressive (AR)** time series techniques

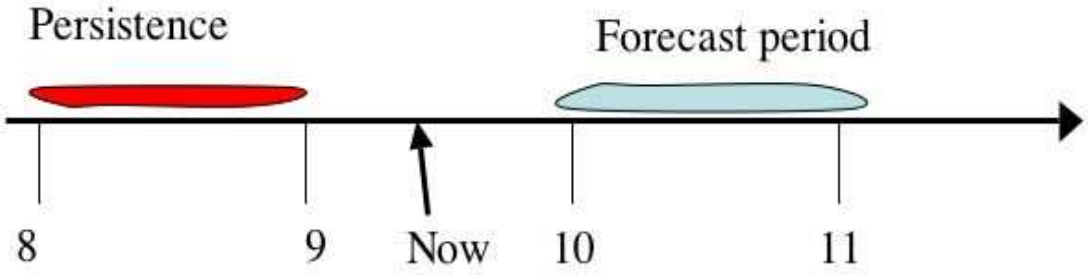
our approach (Gneiting, Larson, Westrick, Genton and Aldrich 2004) is spatio-temporal: **regime-switching space-time (RST)** method

Persistence Forecast

3TIER Environmental FORECAST GROUP

By 9:30 a.m. you must make a prediction for the power produced between 10 a.m. and 11 a.m.

A persistence forecast is using the power produced between 8 a.m. and 9 a.m. (the last hourly power value you have) as your forecast.



Regime-switching space-time (RST) technique

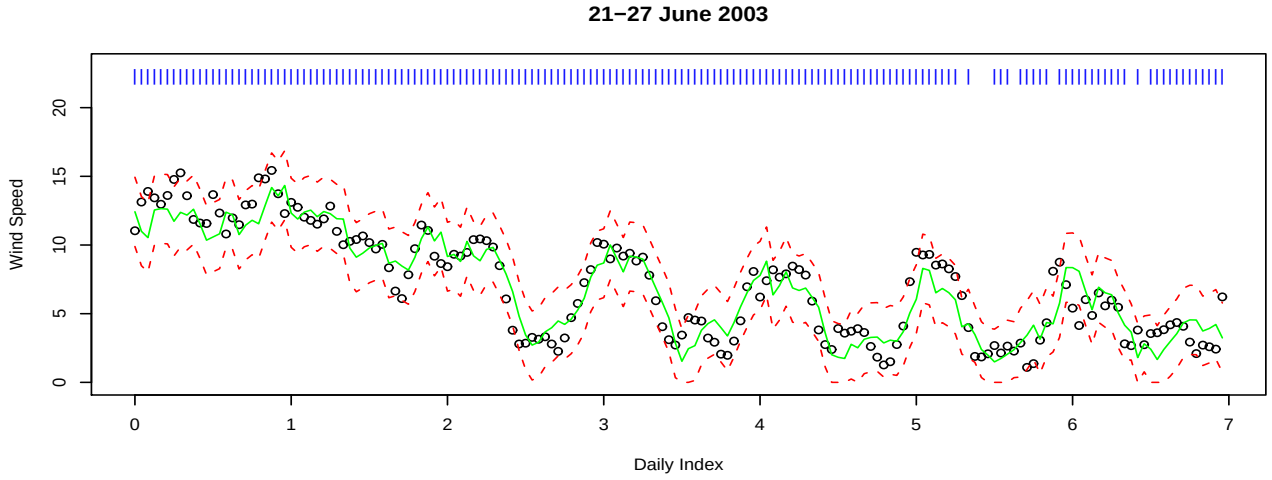
merges meteorological and statistical expertize

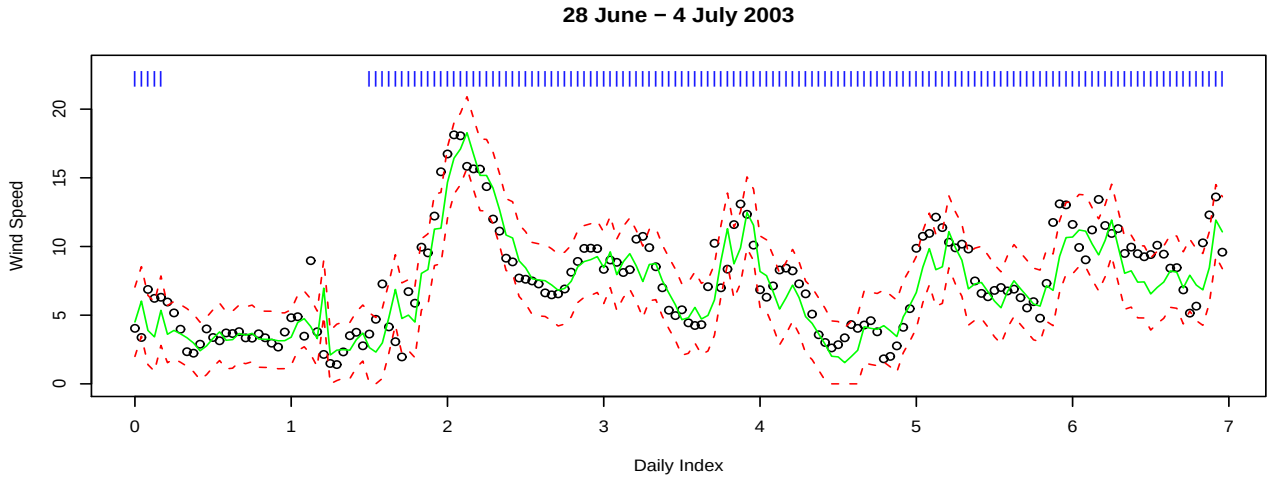
model formulation is parsimonious, yet **takes account of all the salient features of wind speeds:** alternating atmospheric regimes, temporal and spatial autocorrelation, diurnal and seasonal non-stationarity, conditional heteroscedasticity and non-Gaussianity

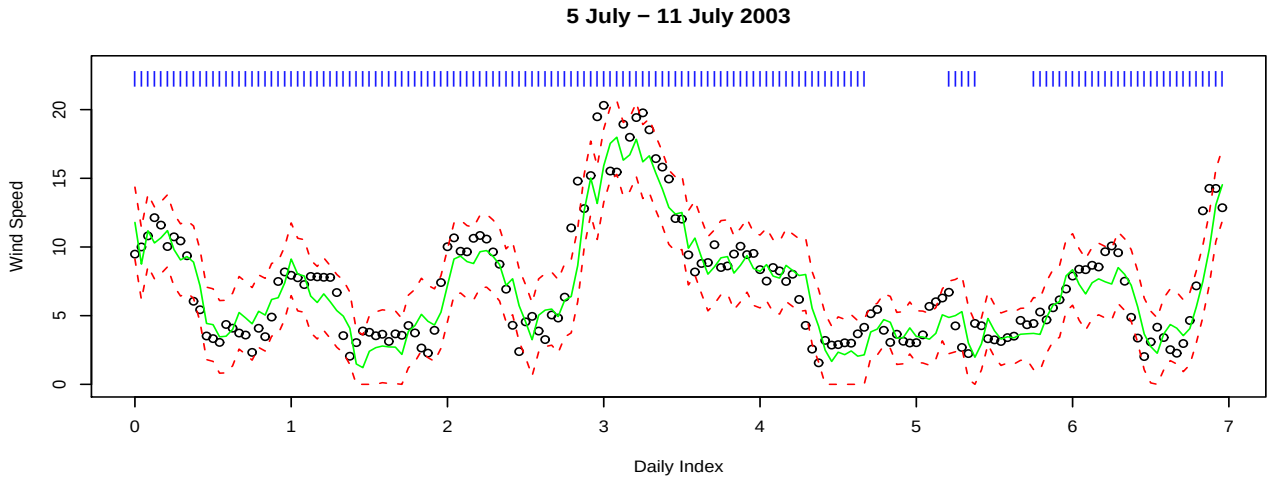
regime-switching: identification of distinct forecast regimes

spatio-temporal: utilizes geographically dispersed meteorological observations in the vicinity of the wind farm

fully probabilistic: provides probabilistic forecasts in the form of predictive CDFs


21–27 June 2003
Wind Speed
0
5
10
15
20
0
1
2
3
4
5
6
7
Daily Index


28 June – 4 July 2003
Wind Speed
0
5
10
15
20
0
1
2
3
4
5
6
7
Daily Index


5 July – 11 July 2003
Wind Speed
0
5
10
15
20
0
1
2
3
4
5
6
7
Daily Index

Verification

evaluation period: May–November 2003

deterministic forecasts: RMSE, MAE

predictive CDFs: PIT histogram, marginal calibration table, CRPS

interval forecasts (90% central prediction interval): coverage, average width, interval score (IntS)

reporting scores month by month allows for significance tests

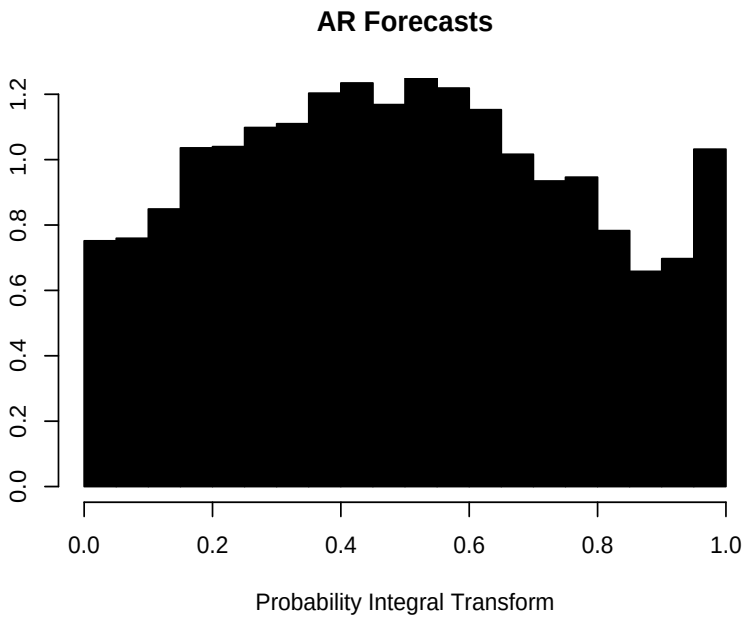
for instance, the RST forecasts had a lower RMSE than the AR forecasts in May, June, . . . , November

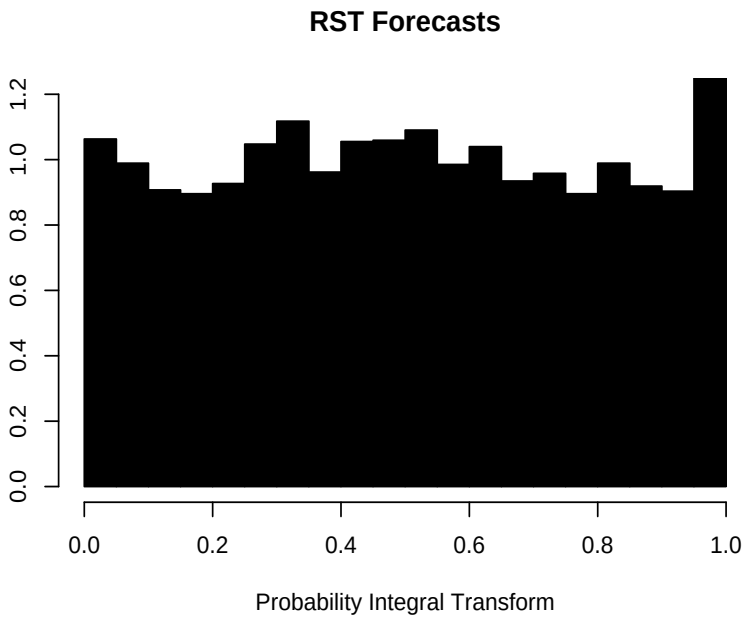
under the null hypothesis of equal skill this will happen with probability $p = \left(\frac{1}{2}\right)^7 = \frac{1}{128}$ only

RMSE ($m \cdot s^{-1}$)	May	Jun	Jul	Aug	Sep	Oct	Nov
Persistence	2.14	1.97	2.37	2.27	2.17	2.38	2.11
AR	2.01	1.85	2.00	2.03	2.03	2.30	2.08
RST	1.75	1.56	1.70	1.78	1.77	2.07	1.88

MAE ($m \cdot s^{-1}$)	May	Jun	Jul	Aug	Sep	Oct	Nov
Persistence	1.60	1.45	1.74	1.68	1.59	1.68	1.51
AR	1.54	1.38	1.50	1.54	1.53	1.67	1.53
RST	1.32	1.18	1.33	1.31	1.36	1.48	1.37

CRPS ($m \cdot s^{-1}$)	May	Jun	Jul	Aug	Sep	Oct	Nov
AR	1.11	1.01	1.10	1.11	1.10	1.22	1.10
RST	0.96	0.85	0.95	0.95	0.97	1.08	1.00


AR Forecasts
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.2
0.4
0.6
0.8
1.0
Probability Integral Transform


RST Forecasts
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.2
0.4
0.6
0.8
1.0
Probability Integral Transform

	5%	50%	95%
Verifications	1.56	6.34	15.62
AR	0.92	6.64	14.95
RST	1.30	6.21	15.12

Cov	May	Jun	Jul	Aug	Sep	Oct	Nov
AR	91.1%	91.7%	89.2%	91.5%	90.6%	87.4%	91.4%
RST	92.1%	89.2%	86.7%	88.3%	87.4%	86.0%	89.0%

Width	May	Jun	Jul	Aug	Sep	Oct	Nov
AR	6.98	6.22	6.21	6.38	6.37	6.40	6.78
RST	5.93	4.83	5.14	5.22	5.15	5.45	5.46

IntS	May	Jun	Jul	Aug	Sep	Oct	Nov
AR	1.74	1.64	1.77	1.75	1.74	2.04	1.86
RST	1.52	1.29	1.41	1.50	1.50	1.83	1.64

Technical reports

www.stat.washington.edu/tilmann

Gneiting, T. and A. E. Raftery (2004)

Strictly proper scoring rules, prediction, and estimation*

Technical Report no. 463, Department of Statistics, University of Washington

Gneiting, T., K. Larson, K. Westrick, M. G. Genton and E. Aldrich (2004)

Calibrated probabilistic forecasting at the Stateline wind energy center: The regime-switching space-time (RST) method

Technical Report no. 464, Department of Statistics, University of Washington

*Introduces scores as positively oriented rewards rather than negatively oriented penalties