Statistical methods for verification of probabilistic forecasts

David Stephenson
Exeter Climate Systems

Plan of this talk

1. Introduction to probability forecasts
2. NAO seasonal hindcast example
3. Signal plus noise model of joint distribution
4. Understanding of verification measures

ECMWF Annual Seminar, 3 Sep 2019, Reading

Доверяй, но проверяй

“Probability does not exist”

Bruno de Finetti
Theory of Probability,
(translation by A Machi and AFM Smith)
2 volumes, New York: Wiley, 1974-5.

Bruno de Finetti 1906-1985

Probability is not an objective reality
- it is a subjective construction
(a model concept for describing data).

It can not be directly observed so can not be directly verified!

But if well-specified, it can be useful for describing data …

Forecast verification aims and approaches

Forecast verification aims to learn about the relationship between observations and forecasts by making use of past performance data. There are three main approaches to forecast verification:

1. **Graphical summary** – visualisation of the raw data is a very sensible first thing to do;

2. **Measures-oriented verification** – use sample statistics of hindcasts and past observations to summarise the (past) performance. Bootstrap resampling can be used to construct confidence intervals.

3. **Distributions-oriented verification** – attempt to statistically model the joint distribution of the observations and forecasts f(y,x). The parameters of this model then provide a complete summary of the forecasting system rather than just a set of rather arbitrary sample statistics. Such a set of statistics would then be sufficient for finding *any* other desired statistic.

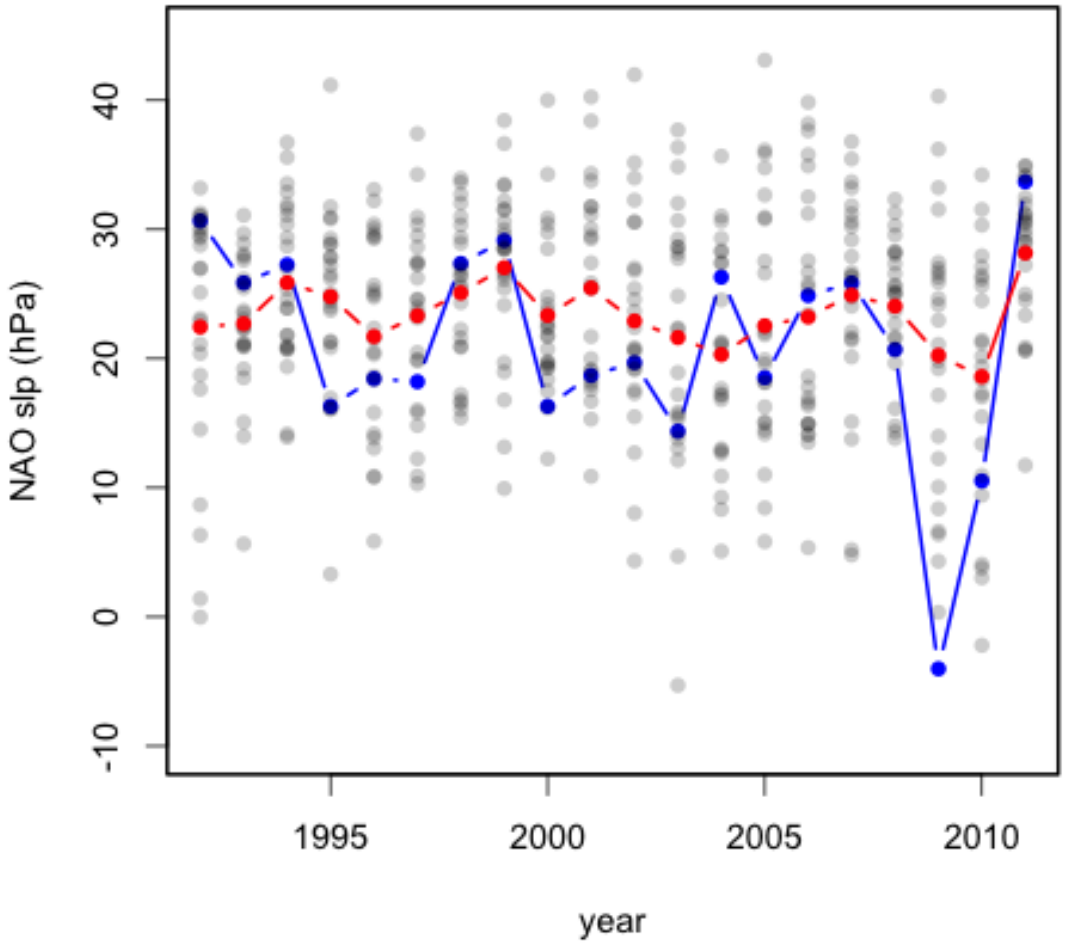
Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. Mon. Wea. Rev., 115, 1330–1338.

North Atlantic Oscillation ensemble hindcast data

GloSea5 ensemble predictions of NAO sea-level pressure:

- t=1,2,…,n Dec-Feb winter observations 1992-2011 $\{y_1,\ldots,y_n\}$
- R=24 hindcasts for each winter $\{x_{t,1},\ldots, x_{t,R}\}$ that can give ensemble mean

Kindly provided by Adam Scaife, Met Office.



Performance measures

Summarise forecast quality by calculating sample statistics (*performance measures*) of previous sets of forecasts (hindcasts). These give guidance on possible quality of future forecasts if one assumes stationarity and ignores sampling uncertainty.

Commonly used performance measures* include:

- Mean Bias (aka Bias in the Mean) (B)
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Correlation
- Ratio of variances of forecasts and observations (aka Bias in Variance)

* Not all of these are either *scores* and/or *metrics* (distances):

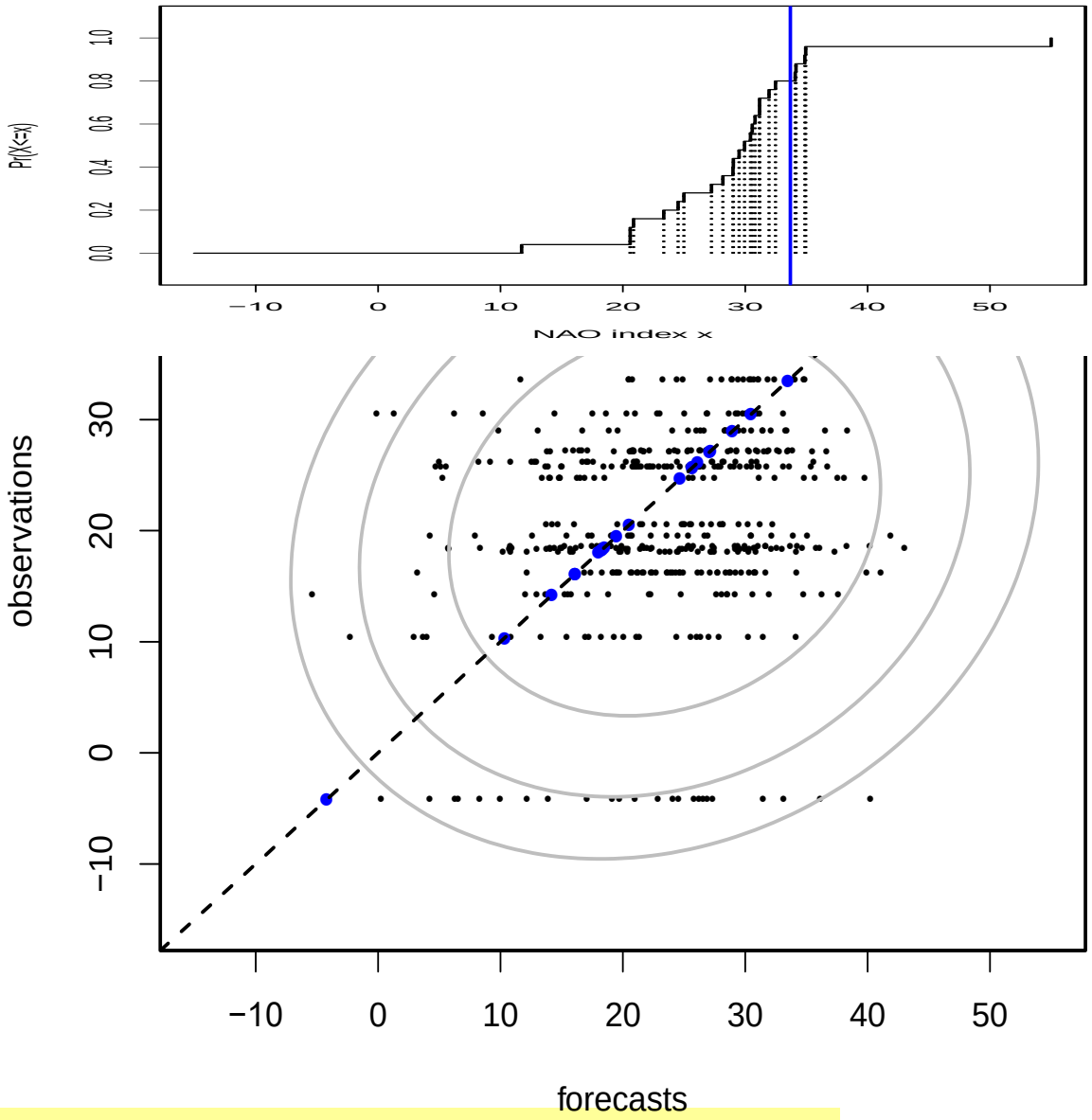
- A *score* is a measure that can be written as the sum of losses over each pair of forecasts and observations (and so is additive over different time periods).
- A *metric* is a distance that is never negative and is only zero when the forecast exactly matches the observation, and satisfies the triangle inequality d(x,z)<= d(x,y)+d(y,z).

Performance of NAO forecasting system

Measure	Definition	Value
Mean Bias	$\frac{1}{N}\sum_{t=1}^{n}(y_t - \bar{x}_t)$	-2.48hPa
Mean Squared Error	$\frac{1}{N}\sum_{t=1}^{n}(y_t - \bar{x}_t)^2$	55.4hPa^2
Mean Absolute Error	$\frac{1}{N}\sum_{t=1}^{n}\lvert y_t - \bar{x}_t \rvert$	5.61hPa
Correlation	$\frac{1}{N s_{\bar{x}} s_y}\sum_{t=1}^{n}(y_t - \bar{y})(x_t - \bar{x})$	0.62
Variance of observations	$s_y^2 = \frac{1}{N-1}\sum_{t=1}^{n}(y_t - \bar{y})^2$	70.65hPa^2
Variance of ensemble mean forecast	$s_{\bar{x}}^2 = \frac{1}{N-1}\sum_{t=1}^{n}(\bar{x}_t - \bar{x})^2$	5.52Pa^2

→ Can more meaningful statistics be found to summarise performance?

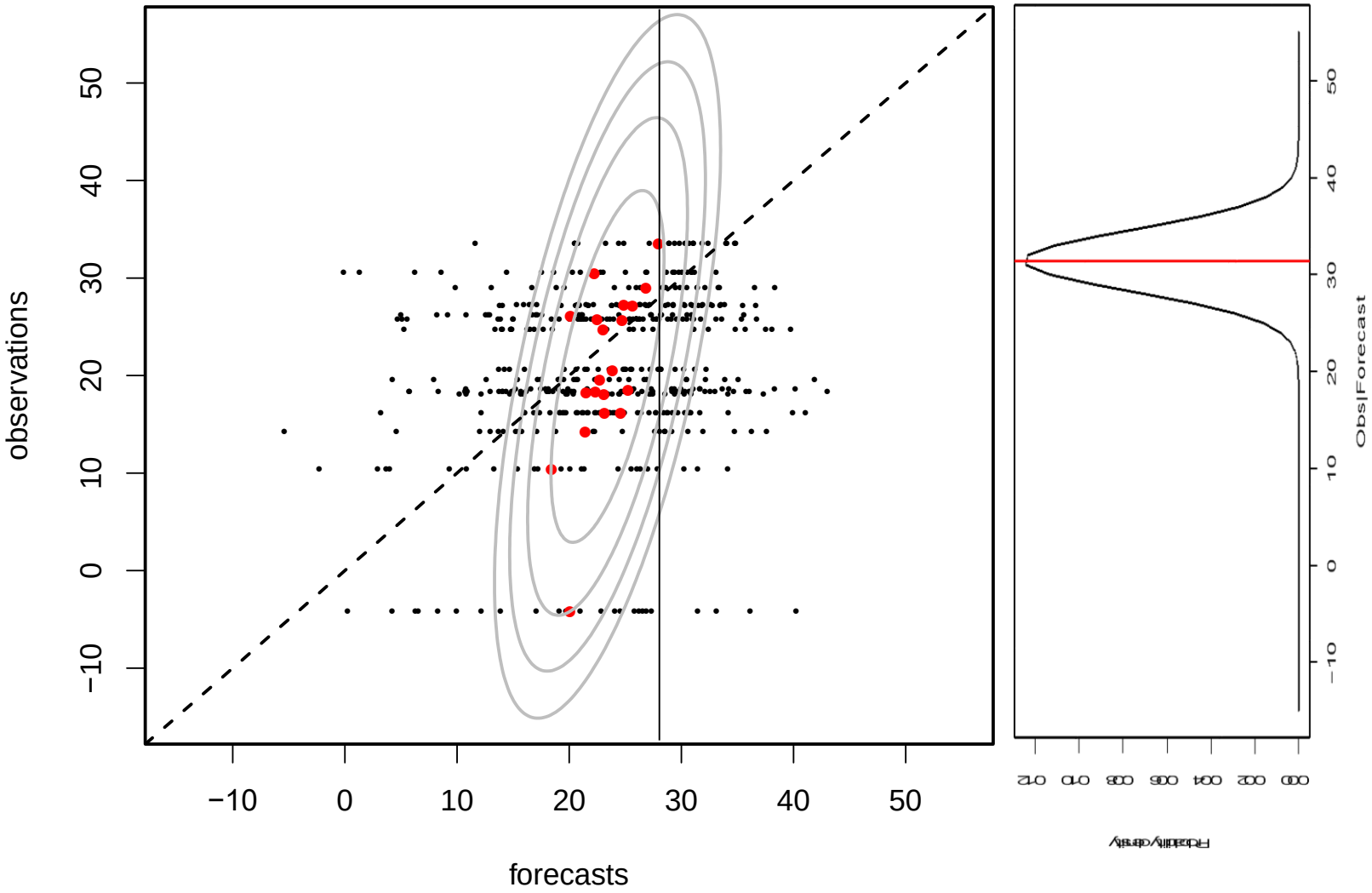
Joint distribution of observations and forecasts



→ Similar spread but not much association

Joint distribution of observation and ens mean

Is there a relationship between the observations and the ensemble mean?



→ Positive association between the ensemble mean and observations

Probability forecasts

Forecasts may be used to construct a predictive probability density function f(y) that an observation could have come from.

If the observations are believed to be *exchangeable* with the forecasts then f(y)=f(x) and the predictive distribution can be estimated directly from the ensemble forecasts in various ways:

- Empirical distributions based on relative frequencies
- Non-parametric smoothing of ensemble forecasts e.g. kernel dressing
- Parametric fits of known distributions to ensemble forecasts

But in general observations are not exchangeable with forecasts and so f(y) is not f(x). The predictive distribution can then be found by either

- Regression modelling of the observations on the forecasts to find f(y|x)
- Modelling the joint density of forecasts and observations f(x,y) and then using Bayes' theorem f(y|x)=f(x,y)/f(x)

Modelling the joint probability distribution

One of the simplest (fewest parameter) models for doing this is the 6 parameter signal-plus-noise model of observations and R-member ensemble forecasts:

$$y_t = \mu_y + s_t + \varepsilon_t$$

$$x_{t,1} = \mu_x + \beta s_t + \eta_{t,1}$$

$$\cdots$$

$$x_{t,R} = \mu_x + \beta s_t + \eta_{t,R}$$

where s, ε, and η are independent normal (Gaussian) random variables with zero mean and constant variances (Siegert et al. 2016). If this is a good model of the system, then performance measures are simply functions of these 6 parameters:

$$\mu_x, \mu_y, \beta, \sigma_s^2, \sigma_\varepsilon^2, \sigma_\eta^2$$

Note: if the data are standardised variables then only 2 parameters are required (and so verification is 2 dimensional e.g. cor(x,y) and cor(x,x) suffice).

Siegert, S., Stephenson, D.B., Sansom, P.G., Scaife, A.A., Eade, R. and Arribas, A., 2016. A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability?. *Journal of Climate*, *29*(3), pp.995-1012.

Estimating model parameters

The simplest approach to find point estimates is to equate sample statistics to what one would expect from the model. For example, the following method of moments approach:

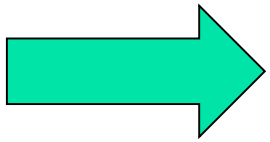
$$v_y = \frac{1}{N-1}\sum_{t=1}^{N}(y_t - \bar{y})^2$$

$$s_{\bar{x}y} = \frac{1}{N-1}\sum_{t=1}^{N}(y_t - \bar{y})(\bar{x}_t - \bar{x})$$

$$c_{rs} = \frac{1}{N-1}\sum_{t=1}^{N}(x_{t,r} - \bar{x}_r)(x_{t,s} - \bar{x}_s)$$

$$c = \frac{1}{R(R-1)}\sum_{r=1}^{R}\sum_{s\neq r}^{}c_{rs}$$

$$v_x = \frac{1}{R}\sum_{r=1}^{R}c_{rr}$$



$$\hat{\mu}_y = \bar{y}$$

$$\hat{\mu}_x = \bar{x}$$

$$\hat{\beta} = cs_{\bar{x}y}^{-1}$$

$$\hat{\sigma}_s^2 = s_{\bar{x}y}\hat{\beta}^{-1}$$

$$\hat{\sigma}_\varepsilon^2 = v_y - \hat{\sigma}_s^2$$

$$\hat{\sigma}_\eta^2 = v_x - \hat{\beta}^2\hat{\sigma}_s^2$$

Note that other estimation methods such as Maximum Likelihood or Bayesian estimation can be used and these can provide information about the uncertainty in the parameter estimates.

Exchangeability

If the model and observations were perfect representations of the real world then one would expect the observations and forecasts to be *exchangeable* i.e. in other words one wouldn't be able to distinguish the observations from being one of the forecasts.

In the signal-plus-noise model, this requires $\mu_x = \mu_y, \sigma_\varepsilon = \sigma_\eta, \beta = 1$

For the NAO hindcasts, we find

$$\hat{\mu}_x = 23.4 \approx \hat{\mu}_y = 20.93$$

$$\hat{\sigma}_\varepsilon = 4.03 << \hat{\sigma}_\eta = 8.20$$

$$\hat{\beta} = 0.22 << 1$$

So the observations are NOT exchangeable with the forecasts because

- the forecasts have much more noise than the observations
- and are much less sensitive to the signal.

Understanding MSE

From the signal-plus-noise model, one expects the MSE of the ensemble mean forecast to be given by:

$$MSE(\overline{x}, y) = (\mu_x - \mu_y)^2 + (1-\beta)^2 \sigma_s^2 + \sigma_\varepsilon^2 + \sigma_\eta^2 / R$$

These terms can be estimated from the model parameters:

5.85 + 31.15 + 15.44 + 2.65

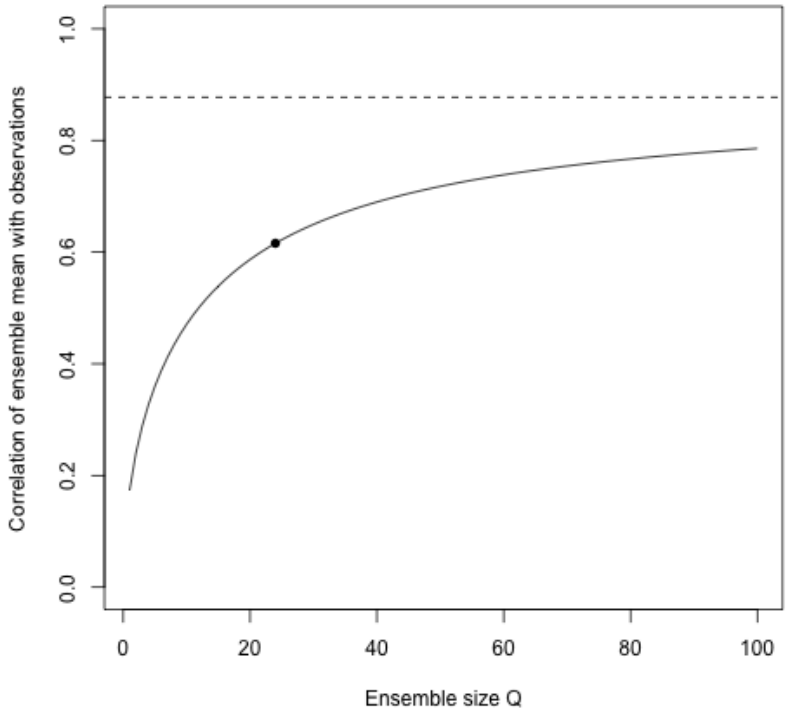
which sum to 55.11, which is close to the value of 55.42 obtained for MSE.

Hence, the biggest contributor to MSE is due to the forecasts having much smaller sensitivity to the signal than the observations (beta << 1). The next major contributor is noise variance in the observations, which can only be reduced by developing forecasting systems having greater signal.

Understanding correlation

From the signal-plus-noise model, one expects the correlation of the ensemble mean forecast with the observations to be given by:

$$cor(\overline{x}_t, y_t) = \beta\sigma_s^2 \left[(\beta^2\sigma_s^2 + \sigma_\eta^2 / R)(\sigma_s^2 + \sigma_\varepsilon^2) \right]^{-1/2}$$



- Useful for designing future ensemble forecasting systems. But note also that the correlation depends strongly on the signal variance so will vary substantially over different verification periods!

Verification of probability forecasts

There are three main complementary approaches:

- **Reliability** – are probabilities well-calibrated? i.e. an event forecast with probability p should occur with relative frequency p.
  - Reliability diagram – plot of mean(y|p) versus p
  - Probability-integral transform (PIT) – if Y~F then $F^{-1}(Y)$~Uniform
  - Rank histogram – special case of PIT where F is the empirical distribution function
- **Probability scores** – loss functions S(p,y) that are minimised by issuing probabilities close to 1 or 0 when observed events occur or don't occur
  - Brier Score
  - Ranked Probability Score
  - Continuous Ranked Probability Score
  - Logarithmic score (ignorance)
  - …
- **Decision-theoretic approaches** – issue warnings/decisions when probabilities exceed decision threshold and then assess the warning classification errors using tools such as Receiver Operating Characteristic curves etc.

Quadratic probability scores

Some widely used quadratic scores:

- Brier Score (BS) – *for probability forecast p=Pr(y=1)*

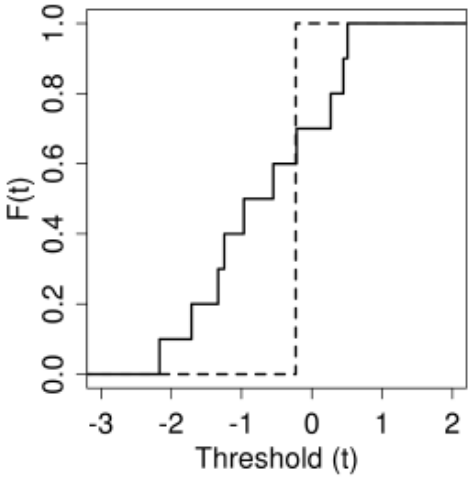
$$S_B(p,y) = (p - y)^2$$

- Ranked Probability Score (RPS) – *for probability forecasts ($p_1,\ldots,p_K$) of ordered categories ($y_1,\ldots,y_K$)*

$$S = \frac{1}{K}\sum_{k=1}^{K} S_B\left(\sum_{j=1}^{k} p_j, \sum_{j=1}^{k} y_j\right)$$

- Continuous Rank Probability Score

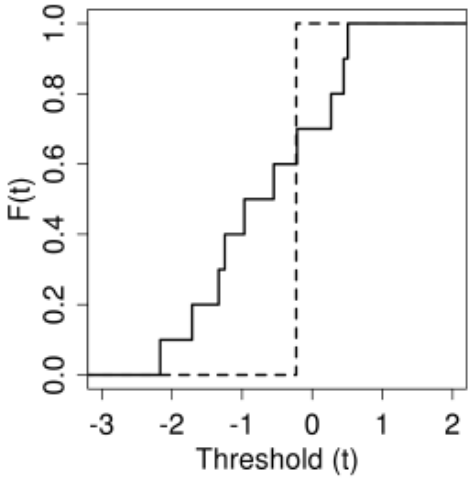
$$S(F,y) = \int S_B(F(t), I(y \le t))\,dt$$



CRPS representation for ensemble

The Continuous Rank Probability Score (CRPS)

$$S(F,y) = \int S_B(F(t), I(y \le t))\,dt$$



can be shown (Gneiting and Raftery 2007) to equal

$$S(F,y) = E_F(|y - X|) - 0.5E_F(|X' - X|)$$

where X and X' are independent draws from the F distribution (i.e. ensemble forecasts).

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102: 359–378.

CRPS for signal-plus-noise model

CRPS can be written as

$$S(F,y) = E_F(|y - X|) - 0.5E_F(|X' - X|)$$

Since for centered Gaussian distributions $X \sim N(0,\sigma^2)$ it can be shown that, $E(X) = \sigma\sqrt{2\pi^{-1}}$ this allows us to write down CRPS for the signal-plus-noise model:

$$S(F,y) = \sqrt{\frac{2}{\pi}\left\{(1-\beta)^2\sigma_s^2 + \sigma_\varepsilon^2 + \sigma_\eta^2\right\}} - \left(1 - \frac{1}{R}\right)\frac{\sigma_\eta}{\sqrt{\pi}}$$

when $\mu_X = \mu_Y$. For ensembles exchangeable with the observations, this gives

$$S(F,y) = \left(1 + \frac{1}{R}\right)\frac{\sigma_\eta}{\sqrt{\pi}}$$

Note that CRPS depends on ensemble size R (see *fair scores*).

Understanding CRPS

The CRPS for equal mean signal-plus-noise model:

$$S(F,y) = \sqrt{\frac{2}{\pi}\left\{\left(1-\beta\right)^2 \sigma_s^2 + \sigma_\varepsilon^2 + \sigma_\eta^2\right\}} - \left(1-\frac{1}{R}\right)\frac{\sigma_\eta}{\sqrt{\pi}}$$

The model parameter estimates give:

4.092 = 8.444 - 4.353 (which are close to data 3.752 = 8.038 - 4.286)

The first term can be decomposed further:

$$\sqrt{\frac{2}{\pi}\left\{\left(1-\beta\right)^2 \sigma_s^2 + \sigma_\varepsilon^2 + \sigma_\eta^2\right\}} = \sqrt{\frac{2}{\pi}\left\{31.1+16.1+64.8\right\}}$$

which shows a sizeable contribution from the signal variance term. This implies that CRPS can vary substantially over different time periods.

Generalised score decomposition

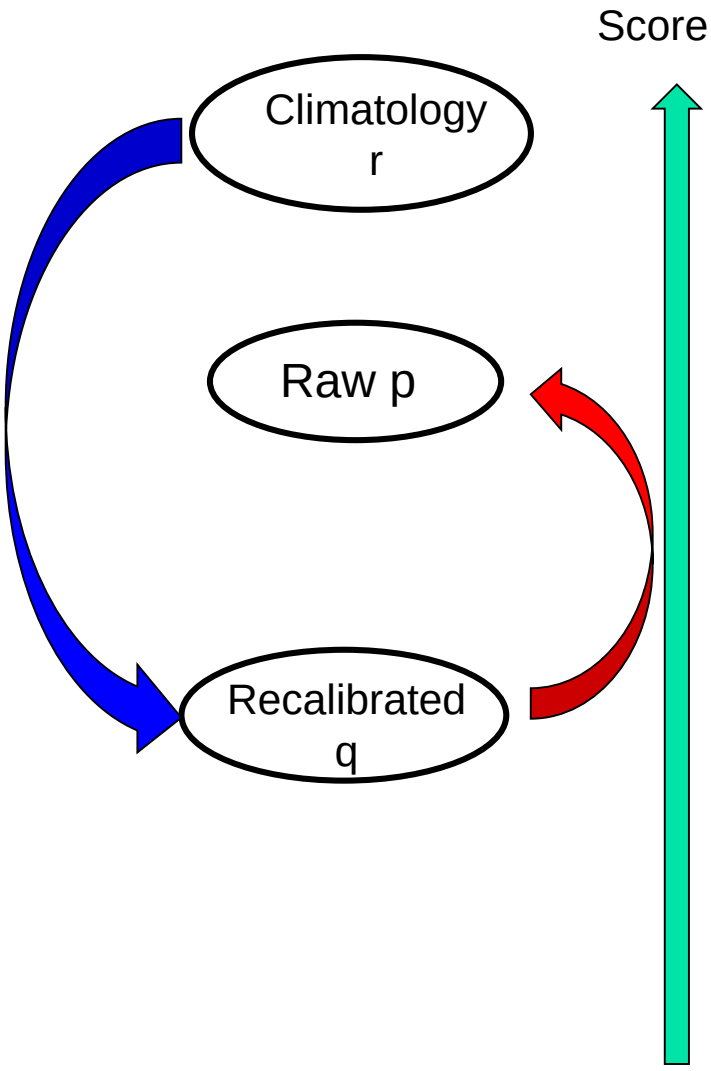
Murphy (1973):
Brier score B(p) = REL – RES + UNC

However calculating RELiability, RESolution and UNCertainty is not easy AND the derivation is complicated and unintuitive.But there is a nice simplification - the components of Murphy's Brier score decomposition are just score differences!

B(p) = [B(p) – B(q)] – [B(r) – B(q)] + [B(r)]
= REL – RES + UNC

where p are the raw forecasts with Brier score B(p), q are recalibrated forecasts with Brier score B(q) and r are (constant) climatological forecasts with score B(r)

The derivation generalises trivially to different verification scores (e.g. CRPS), different recalibration methods (e.g. regression techniques), and different reference forecasts (e.g. persistence).



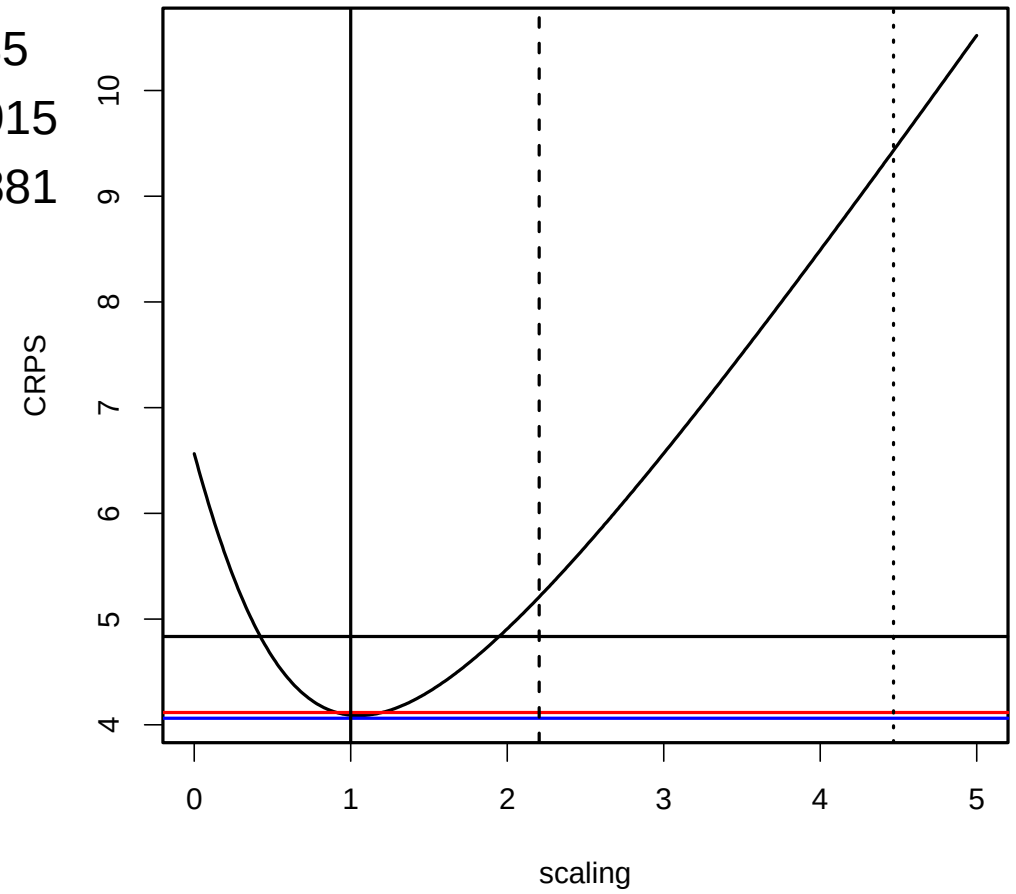
Siegert S. (2017) Simplifying and generalising Murphy's Brier score decomposition, Quarterly Journal of the Royal Meteorological Society, volume 143, no. 703, pages 1178-1183

How much improvement by recalibration?

If we multiply the forecasts by a scaling factor how much can we improve CRPS?
Dotted line = 1/beta Dashed line = regression slope of y on x

Climatology CRPS = 4.835
Raw forecast CRPS = 4.0915
Best forecast CRPS = 4.0881

- REL=0.0035
- RES=0.747
- UNC=4.835



- Linear recalibration does not improve CRPS for these forecasts
  (RELiability component is small yet probability forecasts are not well-calibrated!).

Conclusions

- Descriptive (measures-oriented) verification is a useful starting point but much more insight can be gained by modelling the forecast-observation joint distribution;

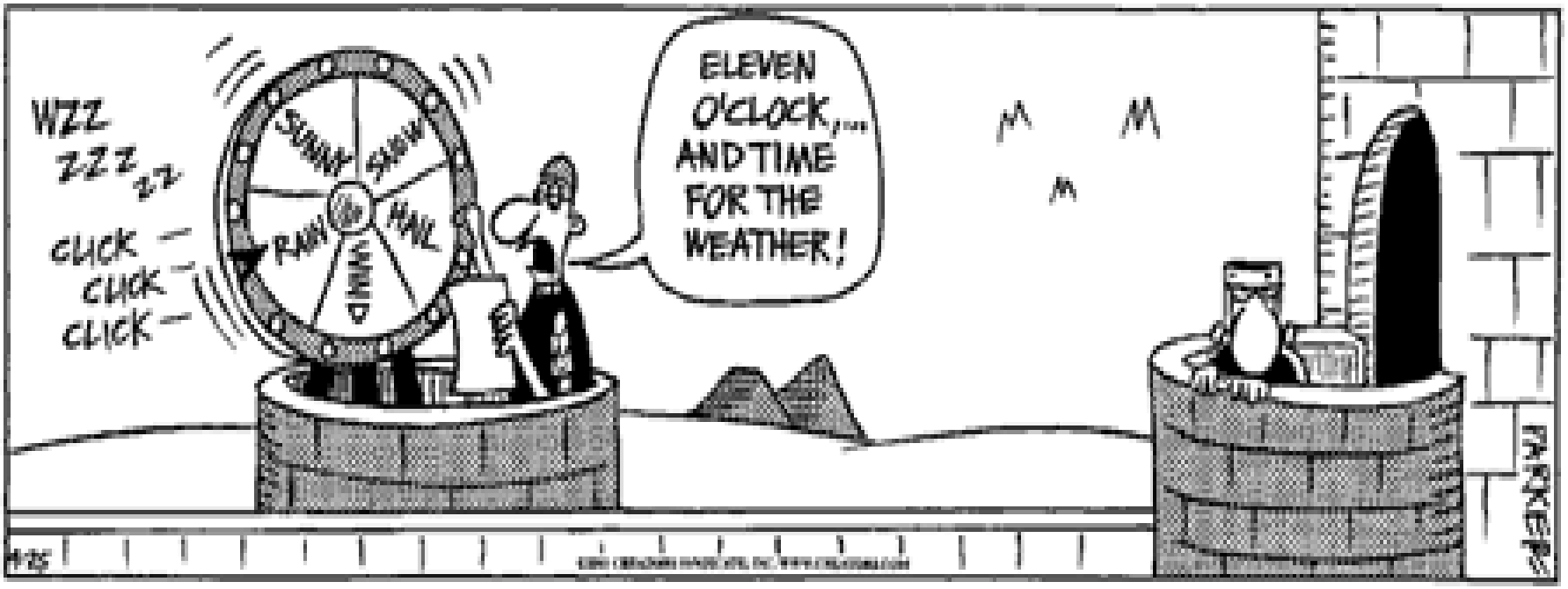
- The signal-plus-noise model provides a useful framework for understanding seasonal and longer term forecasting systems;

- It can be used to estimate reliability, sources of MSE, CRPS and its decompositions, and predict how measures will depend on ensemble size;

- Most of the usual summary measures involve more than one of the 6 model parameters and so can lead to confusion if used to infer properties such as reliability, overdispersion etc.

- Sampling uncertainty is a major issue for longer range forecasts where the sample size of observations is generally small. It therefore makes sense to exploit any prior beliefs using Bayesian estimation approaches such as MCMC (see Siegert et al. 2016).

Thank you for your attention



Any questions?

d.b.stephenson@exeter.ac.uk

UNIVERSITY OF EXETER

References

Siegert S, Stephenson DB, Sansom PG, Scaife AA, Eade R, Arribas A. (2016) A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability?, *Journal of Climate*, volume 29, no. 3, pages 995-1012.

Siegert S, Stephenson D. (2018) Forecast recalibration and multi-model combination, *Sub-seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting*, Elsevier. Editors: Andrew Robertson & Frederic Vitart, 585 pages.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102: 359–378.

Siegert S. (2017) Simplifying and generalising Murphy's Brier score decomposition, Quarterly Journal of the Royal Meteorological Society, volume 143, no. 703, pages 1178-1183, DOI:10.1002/qj.2985.

Useful verification software in R

https://cran.r-project.org/web/packages/

SpecsVerification	Forecast verification routines developed at Exeter
s2dverification	Set of Common Tools for Model Diagnostics
easyVerification	Verification for large ensemble data sets

verification	Weather Forecast Verification Utilities
SpatialVx	Spatial Forecast Verification
ternvis	Visualisation and verification of ternary probabilistic forecasts

Uncertainty quantification

Performance measures are calculated using finite samples of forecasts at finite number of past times. This gives rise to two sources of sampling uncertainty which can be quantified using approaches such as:

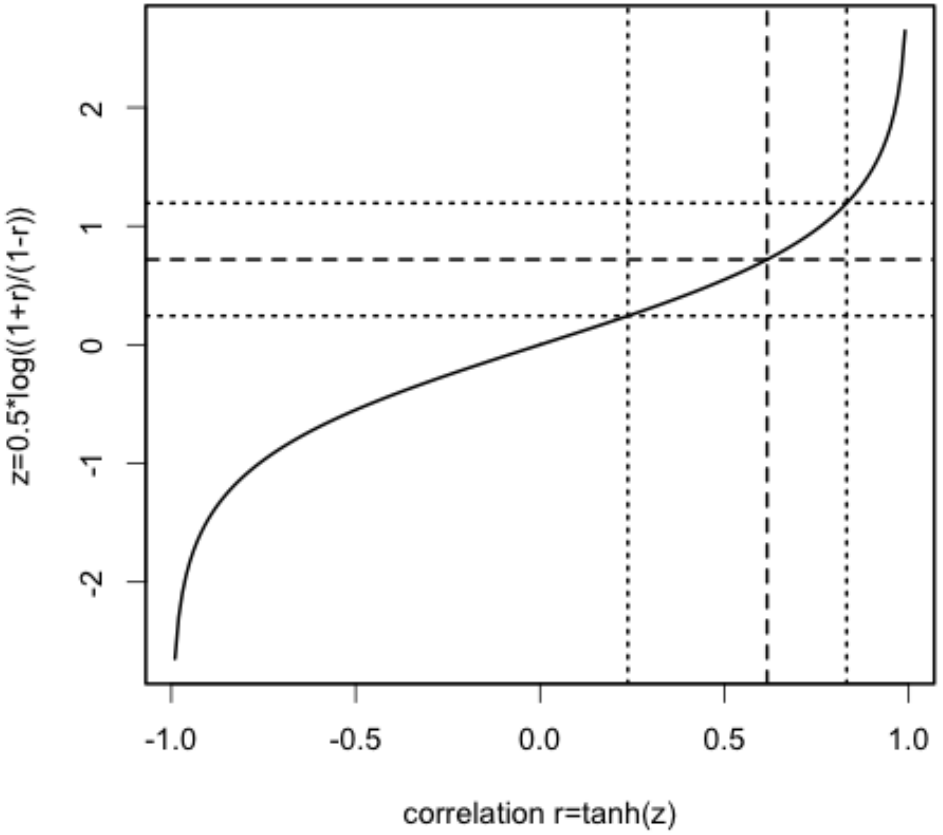
- **Large-sample (asymptotic) theory** – it is sometimes possible to know the approximate sampling distribution for a measure in the limit of large sample size e.g. correlation (see next slides)

- **Bootstrap resampling** – create a new sample of "data" by resampling the original data *with replacement* and then recalculate the measure. Repeat this many (e.g. 1000) times to get a sample of measure values. Either forecasts in a year and/or different years can be resampled depending on the question of interest. Note – resampling relies on assuming independence of the data values.

- **Model simulation (parametric bootstrap)** – fit a statistical model to the data and then use this model with new random numbers to generate new artificial (synthetic) data sets. Parameter uncertainty can be accounted for by using Bayesian methods such as Markov Chain Monte Carlo (Siegert et al. 2016)

Uncertainty in correlation

For large samples of normally distributed data, Fisher (1915) showed that a non-linear transformation of correlation Z(r) is normally distributed with variance 1/(N-3). This is useful for finding confidence intervals on correlation:



→ Using a +/-1.96$(N-3)^{-0.5}$ interval for Z gives a 95% confidence interval for correlation of (0.24,0.83). Does not include zero so skill is significant at 5% level.