Some additional, “not widely-used”, skill scores in climate science

Nachiketa Acharya

npa5302@psu.edu

Center for Earth System Modeling, Analysis, and Data,

Department of Meteorology and Atmospheric Science, Pennsylvania State University, USA

Topics:

This lecture presents some scores which are not widely-used in climate science, but which nevertheless provide complementary insights into the forecast performance, hence they could be added to the commonly computed scores. These will include:

- Index of agreement.
- Forecast comparison using random walk.

- Ignorance skill score
- AIC/BIC
- Point biserial correlation

Index of Agreement

- The relative difference measures such as the ratio between RMSE and observed climatology have the limitation that they are not bounded and are unstable for very small (near zero) climatology of observation.

- As a remedy, Willmott (1982) proposed new skill metrics called ‘index of agreement (d)’, and it is bounded by 0 and 1 ($0 \leq d \leq 1$).

$$d_{orig} = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(|P_i - \bar{O}| + |O_i - \bar{O}|)^2}$$

Index of Agreement: Example

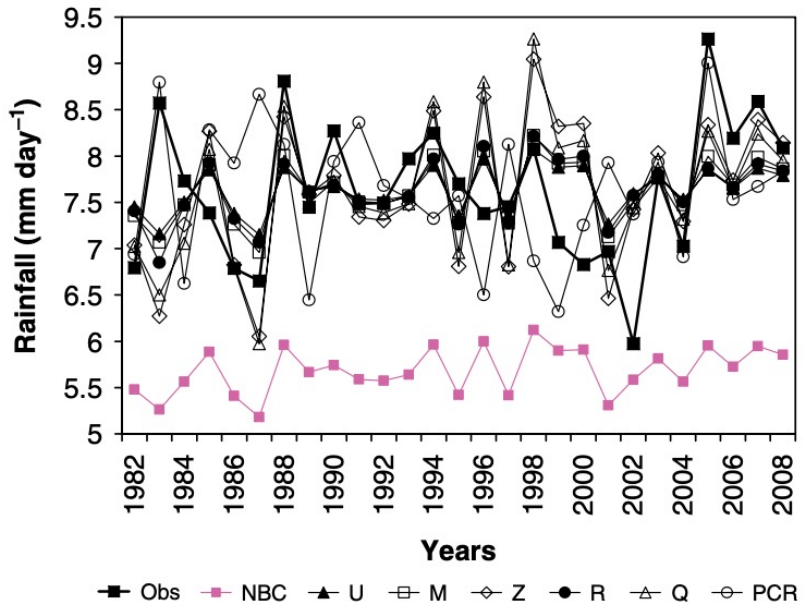


Figure 1. Interannual variability of rainfall from observation, raw model and six bias correction methods for June–July–August–September

Statistic	Obs	NBC	U	M	Z	R	Q	PCR
Mean (mm day^{-1})	7.63	5.68	7.63	7.63	7.63	7.63	7.61	7.61
SD (mm day^{-1})	0.75	0.25	0.25	0.34	0.78	0.34	0.76	0.72
RMSE (mm day^{-1})	–	2.06	0.69	0.44	0.83	0.72	0.80	0.83
Index of agreement (d)	–	0.37	0.45	0.58	0.65	0.48	0.66	0.60

Acharya et al, (2012).

Index of Agreement

- Willmott et al. (1985) noted that the use of squared differences in its calculation algorithm might result in high values of this index ($d_{orig} \approx 1$) even in the presence of large errors. In addition, sums-of-squares-based measurements vary in response to both variability and central tendency within a set of deviations.

- A modification in the index that replaces the square function by the modulus of the deviations.

$$d_{mod} = 1 - \frac{\sum_{i=1}^{n} |P_i - O_i|}{\sum_{i=1}^{n} (|P_i - \bar{O}| + |O_i - \bar{O}|)}$$

- The advantages of d_{mod} over d_{orig} is that errors and differences are given their appropriate weighting factors.

- d_{mod} may be regarded as a more rigorous method than d_{orig} as d_{mod} tends to approach its maximum value more slowly than the d_{orig} as the predicted values approach the observed data.

Index of Agreement

- Willmott et al. (2012) stated that the overall range of d_{orig} and d_{mod} [0:1] is narrow to adequately represent the great variety of forms that predicted values can diff er from observed data.
- Therefore, these authors proposed a new index, referred to as the refined index of agreement (d_{ref}), that is bounded by –1.0 and 1.0. Willmott et al. (2012) claimed that the d_{ref} is more rationally related to model accuracy than d_{orig} and d_{mod}

$$d_{ref} = \begin{cases} 1 - \frac{\sum_{i=1}^{n}|P_i - O_i|}{2\sum_{i=1}^{n}|O_i - \bar{O}|}, \text{when} \\ \sum_{i=1}^{n}|P_i - O_i| \leq c \ \sum_{i=1}^{n}|O_i - \bar{O}| \\ \frac{2\sum_{i=1}^{n}|O_i - \bar{O}|}{\sum_{i=1}^{n}|P_i - O_i|} - 1, when \\ \sum_{i=1}^{n}|P_i - O_i| > 2 \ \sum_{i=1}^{n}|O_i - \bar{O}| \end{cases}$$

Forecast comparison using Random walk

- Is one forecast model better than another?
- For example, suppose you have a criterion for deciding which model makes a “better” forecast of an event. Then, if two models are equally skillful, the probability that one model beats the other is 50%.
- The problem looks exactly like the problem of deciding if a coin is fair. It can be test if a coin is fair by flipping it several times and checking that about half of the tosses are heads. More precisely, the number of heads follows a binomial distribution with p = 1/2, and the null hypothesis is rejected if the number of heads falls outside the 95% interval computed from that distribution.
- Similarly, to compare the skill of models A and B, one should simply count the number of times model A produces a better forecast than B, and compare that count to the 95% interval.

DelSole and Tippett, (2016).

Sign test

- The Sign test, a nonparametric test, is a statistical method to test for consistent differences between pairs of observations.
- This is a good measurement because it can be evaluated simply by computing differences in some measure and then counting the number of positive signs.

Let $p = \Pr(X > Y)$, and then test the null hypothesis H_0: $p = 0.50$. In other words, the null hypothesis states that given a random pair of measurements (x_i, y_i), then x_i and y_i are equally likely to be larger than the other.

To test the null hypothesis, independent pairs of sample data are collected from the populations $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. Pairs are omitted for which there is no difference so that there is a possibility of a reduced sample of m pairs.[4]

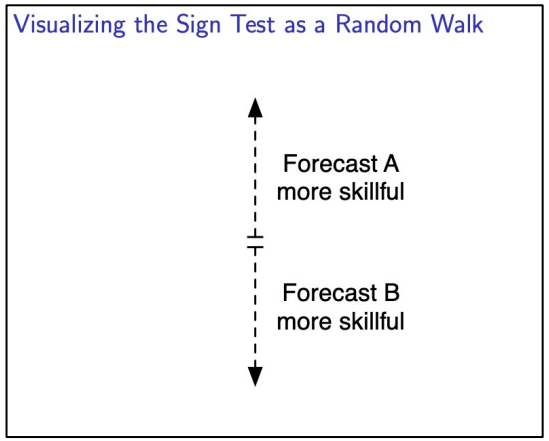
Then let W be the number of pairs for which $y_i - x_i > 0$. Assuming that H_0 is true, then W follows a binomial distribution $W \sim b(m, 0.5)$.

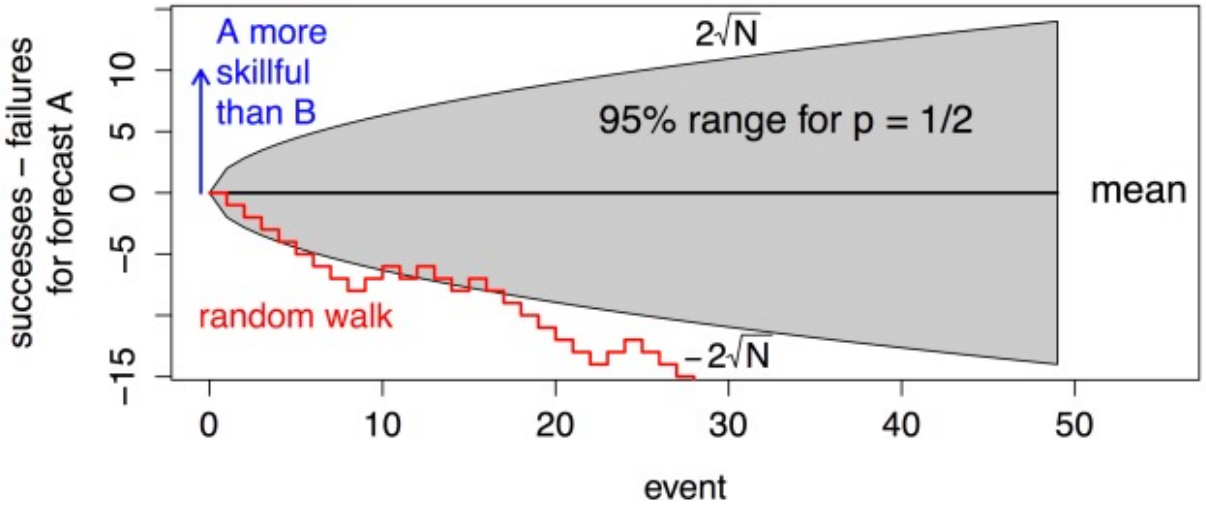
Since the test statistic is expected to follow a binomial distribution, the standard binomial test is used to calculate significance. The normal approximation to the binomial distribution can be used for large sample sizes, $m > 25$.[4]

Wikipedia

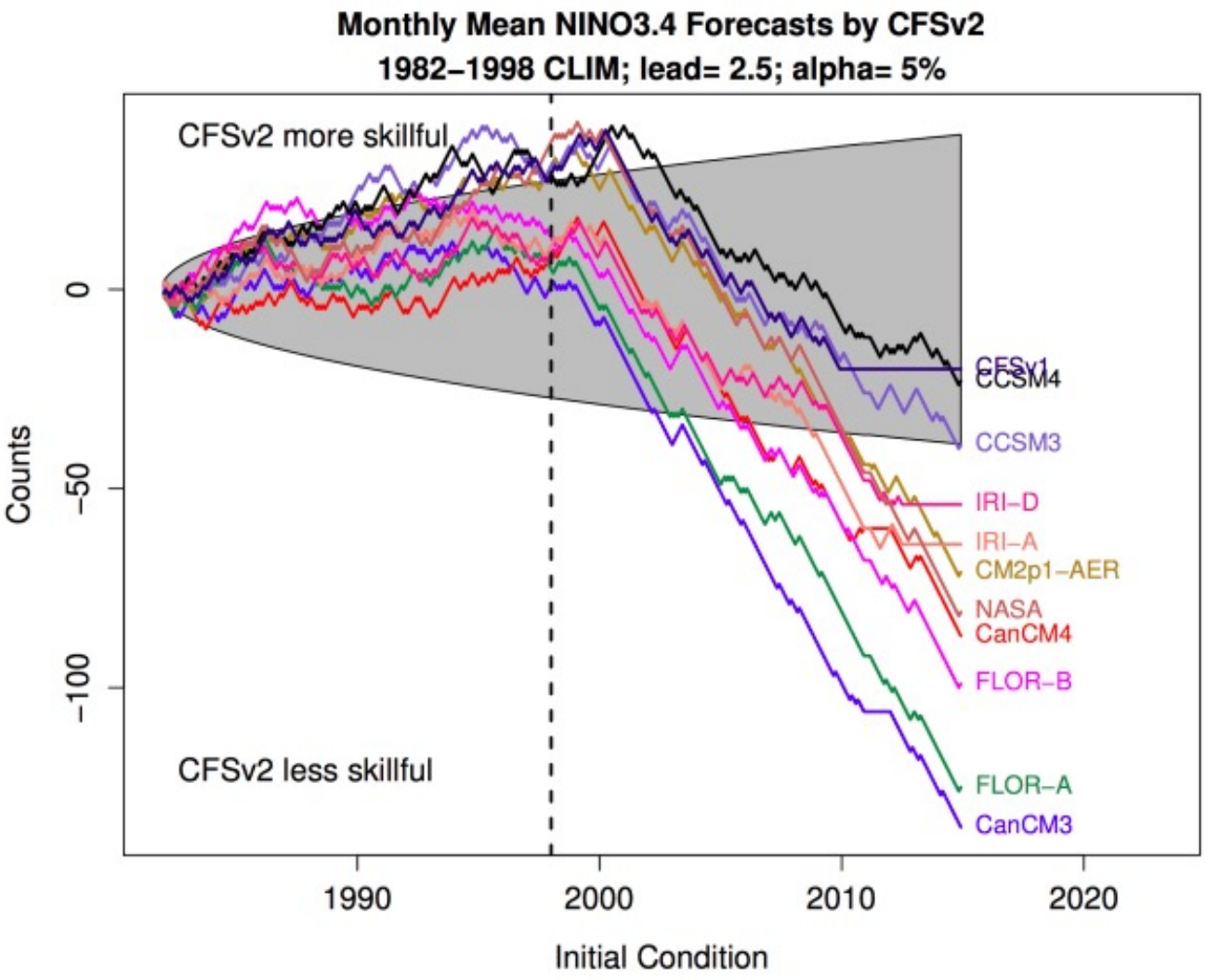
Forecast comparison using Random Walk

- An informative way to summarize the results of the sign test is to display a sequence of sign tests in the form of a random walk.
- At each event, the particle moves up one unit if forecast A is more skillful than B, otherwise it moves down one unit. If the forecasts are equally skillful, then the upward and downward steps are equally probable, and the average location is y = 0.
- the 95% interval for a random walk has a very simple form: it is ±2√N, where N is the number of independent forecast events. The hypothesis that the forecasts are equally skillful is rejected when the particle walks outside the 95% interval.





Forecast comparison using Random Walk :Example



- Random walk test comparing monthly mean forecasts of the NINO3.4 index at 2.5-month lead between CFSv2 and other models in the NMME dataset.
- From 1982 to about 2000, the random walks drifted upward, with some models going above the 95% interval, indicating that CVSv2 was more skillful than these models.
- However, after 2000, there is an abrupt change in skill, so by the end of the period most models lie below the 95% interval, indicating that these models were more skillful than CFSv2.

DelSole and Tippett, (2016).

Logarithmic skill score (LSS)

- Given the forecast, how much additional information is needed to determine what the verifying categories were?
- The logarithmic score (LS) is an alternative to the RPS and for an individual forecast is the simply the logarithm of the forecast probability of the observed category.
- The LS does not depend on the forecast probabilities of the other categories, and this property means that the LS is a local score, unlike RPS.
- If the observations falls in the i-th category, then the LS of the forecast is log(p_i).
- logarithmic skill score (LSS) defined as $\mathrm{LSS} = \overline{\mathrm{LS}} - \overline{\mathrm{LS}}_{\mathrm{ref}}\,,$
- where LS and LS_{ref} are the LS of the forecast being evaluated and a reference forecast, respectively, averaged over multiple forecasts.

Ignorance Score (IGN) and Skill Score

- A scoring rule, based on “relative entropy”, (Roulston & Smith, 2002).
- It measures the information deficit, or ignorance, of a person having a probabilistic forecast but not knowing the actual outcome.
- Units are bits of information.
- Each bit of ignorance represents a factor-of-2 increase in uncertainty.
- Due to its relationship to Shannon’s information entropy, IGN is frequently used as a proxy for forecast utility, or the amount of information gain expected from a forecast

Entropy (Shannon Entropy)

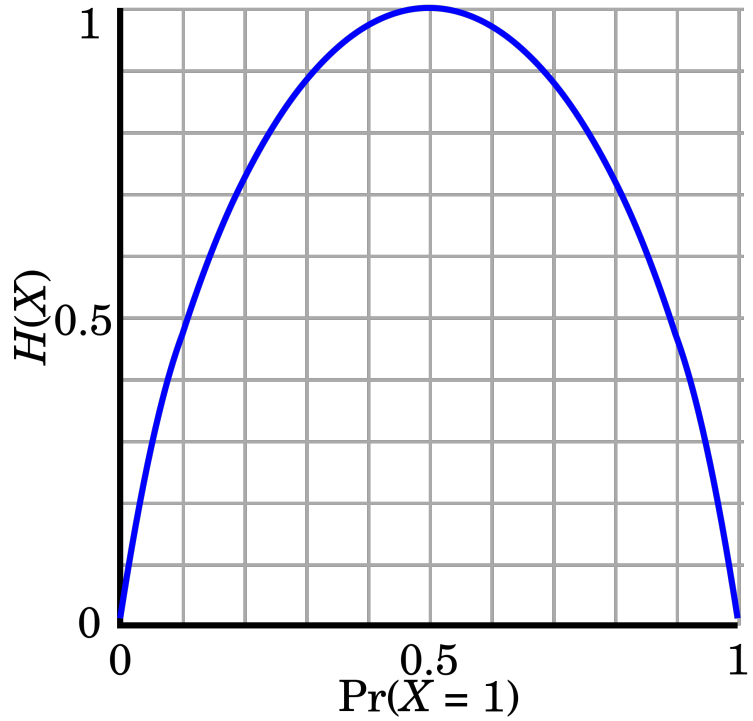
In information theory, the entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent in the variable's possible outcomes.

Given a discrete random variable X, with possible outcomes $x_1, \ldots, x_n$, which occur with probability $\mathrm{P}(x_1), \ldots, \mathrm{P}(x_n)$, the entropy of X is formally defined as:

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log \mathrm{P}(x_i)$$

Entropy is the number of bit to store an information

- Consider tossing a coin with known, not necessarily fair, probabilities of coming up heads or tails.
- The entropy of the unknown result of the next toss of the coin is maximized if the coin is fair (that is, if heads and tails both have equal probability 1/2).
- To communicate the outcome of a coin flip (2 possible values) will require an average of at most 1 bit of useful information (exactly 1 bit for a fair coin).
- To find the number of bits of useful information by computing the Binary logarithm for the uncertainty reduction factor

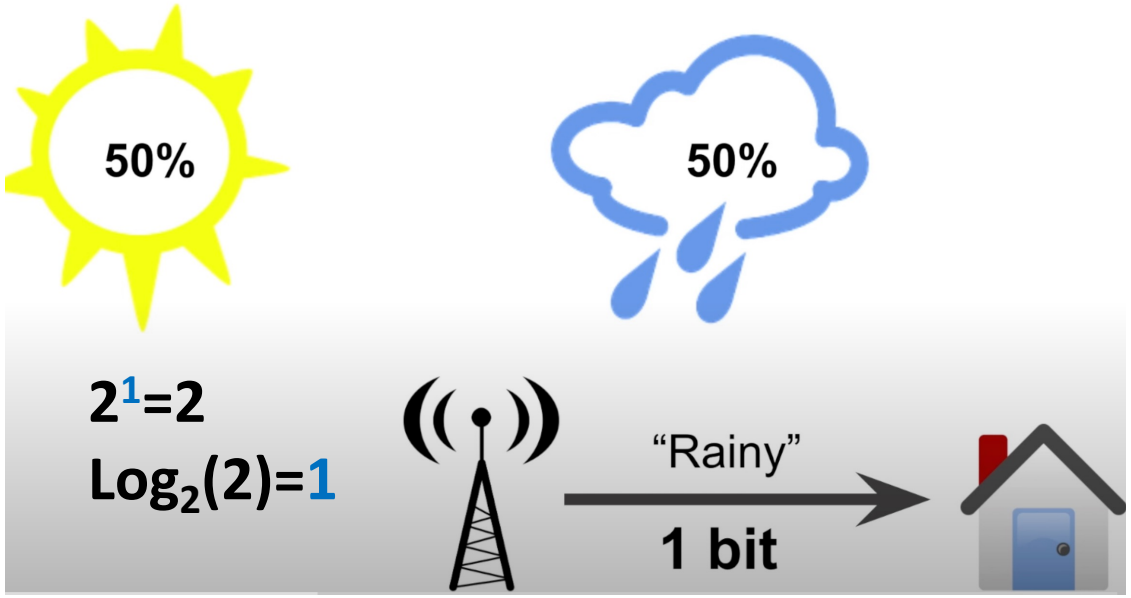


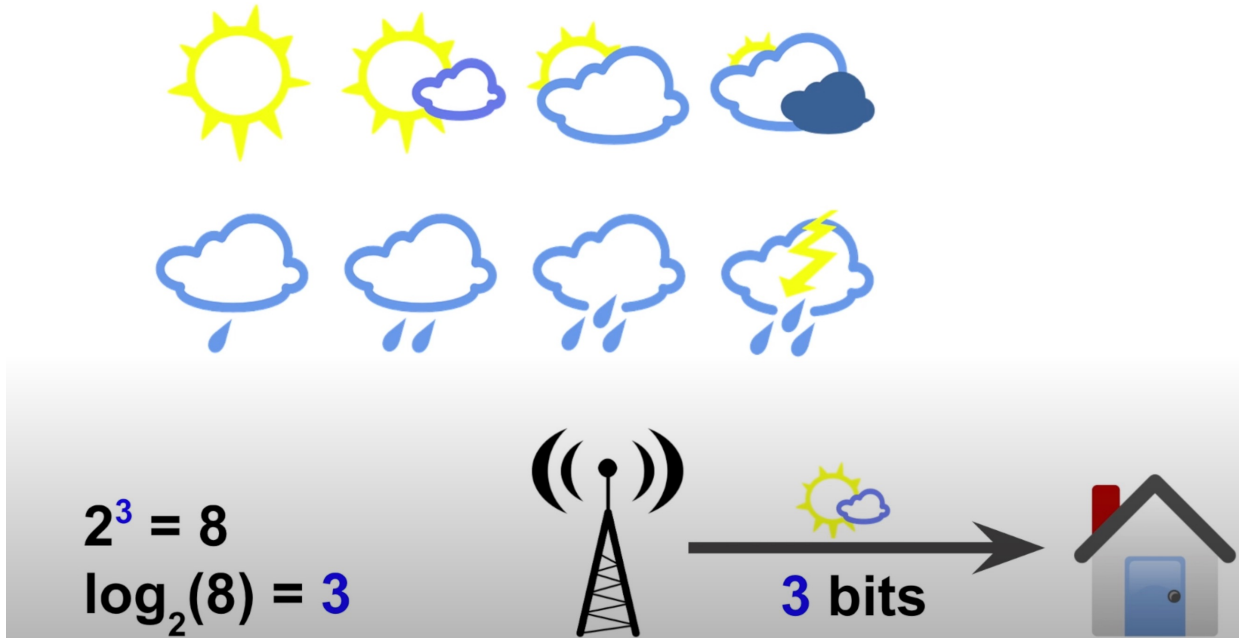
$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i)$$

$$= -\sum_{i=1}^{2} \frac{1}{2} \log_2 \frac{1}{2}$$

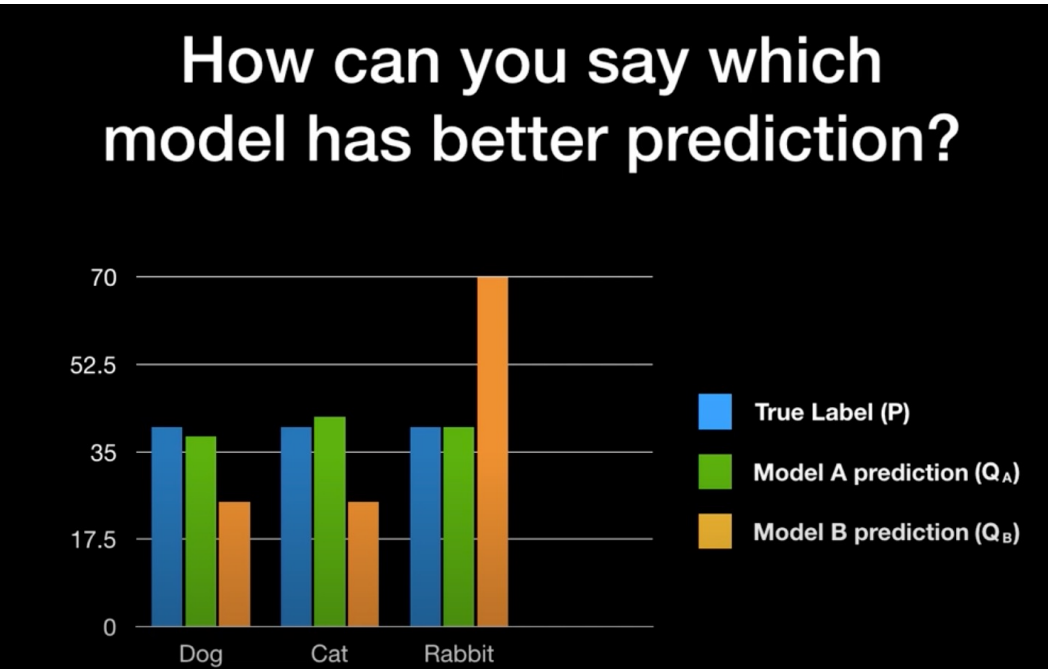
$$= -\sum_{i=1}^{2} \frac{1}{2} \cdot (-1) = 1$$

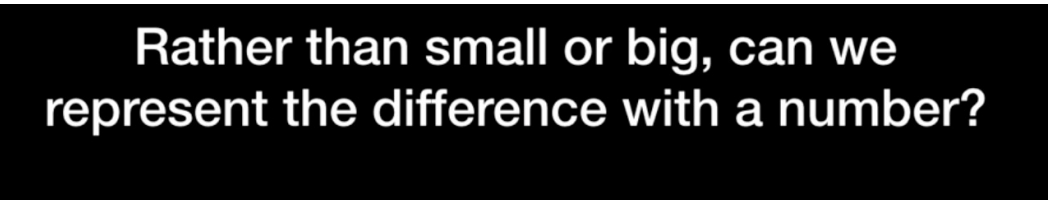
Entropy is the number of bit to store an information





Realtive Entropy (Kullback-Leibler Divergence)





How much Q's are relatively differing from P?

Relative Entropy D(P||Q) is often called
the information gain achieved if P would be used
instead of Q which is currently used.

D(P||Q) can be constructed by measuring the
expected number of extra bits required
to code samples from P using a code optimized
for Q ather than the code optimized for P.

D(P||P)=0 ; no information gain
$D(P||Q_A)= 0.25$; Q_A is similar to true (P) distribution
$D(P||Q_B)= 1.85$

Ignorance Score and Skill Score

The Ignorance Score (IGN), or negative log-likelihood score, of a probabilistic forecast of *n* categories can be written as (Good 1952; Roulston & Smith, 2002):

$$IGN = -\log_2(p_k) \qquad k = 1..n$$

and it can be decomposed into reliability, resolution and uncertainty terms:

$$IGN = REL - RES + UNC$$

(Weijs et al., 2010; Wilks, 2018)

calibration sharpness obs distribution

$$ISS = -\frac{\log_2(p_k)}{\log_2(n)}$$

- > 1 Less info than climatology
- $= 1$ As good as climatology
- < 1 More info than climatology

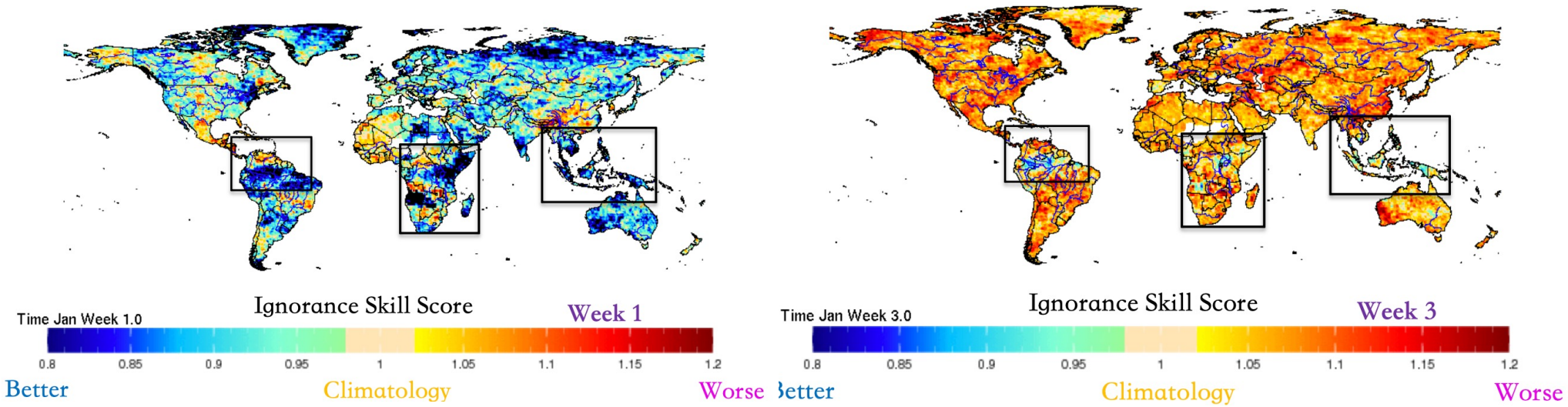
where *n* is the number of categories in the probabilistic forecast.

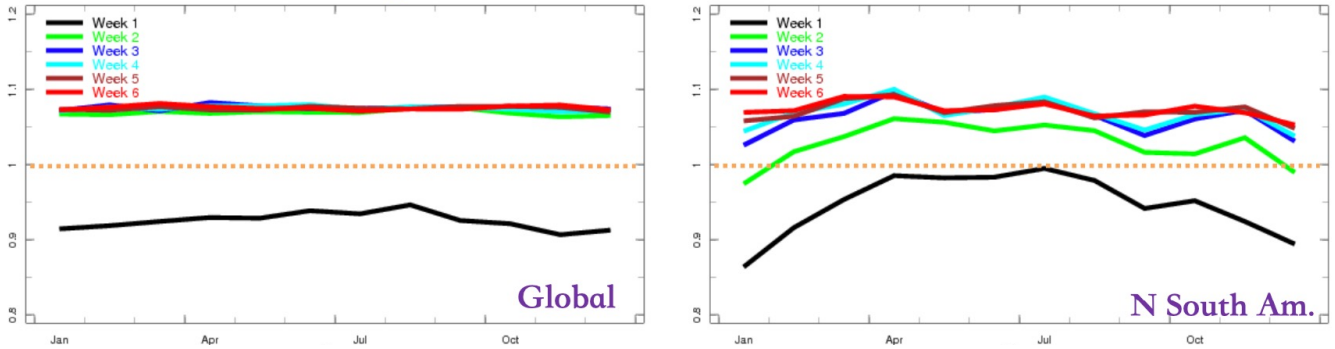
- The ignorance skill score is negatively oriented and cannot be negative and smaller values of the score are better.
- A perfect forecast has zero IGN and ISS.

Exmple: Ignorance Score and Skill Score

Skill Assessment

- Model: ECMWF
- Rainfall
- Probabilistic Hindcasts
- Obs: CPC Unified
- All initializations available per month (8-9)
- **Uncalibrated**
- IGN, RPSS, Brier and decompositions, for Week 1-6





Courtesy: A. Muñoz (IRI)

Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC)

Akaike information criterion (AIC)

$AIC = -2 (\ln (\text{likelihood})) + 2 K$

Bayesian information criterion (BIC)

$BIC = -2 \times \ln(\text{likelihood}) + \ln(N) \times k$

K=parameter

N=sample size

- AIC and BIC are ways for probabilistic model selection.
- Akaike and Bayesian Information Criterion are two ways of scoring a model based on its log-likelihood and complexity.
- The AIC can be termed as a measure of the goodness of fit of any estimated statistical model.
- The BIC is a type of model selection among a class of parametric models with different numbers of parameters.
- Penalty for additional parameters is more in BIC than AIC
- It might be best to use AIC and BIC together in model selection. Less value of both is better model.

Example:

- Suppose we make probabilistic forecast for RX1 (Max 1-day annual precipitation) and R95p (Annual total precipitation when RR >95th percentile) using GEV with different cases.

Generalized Extreme Value (GEV) distribution function

$$\Pr\{M_n \le z\} = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

$\mu \in \mathbb{R}$ (location parameter)

$\sigma > 0$ (scale parameter)

$\xi \in \mathbb{R}$ (shape parameter)

Look familiar?

Defined where the part inside the [ ] is positive.

- Model 1: GEV($\mu_t=\mu$, $\sigma_t=\sigma$, $\xi_t=\xi$)
- Model 2: GEV($\mu_t=\mu_0+\beta_t$, $\sigma_t=\sigma$, ξ)
- Model 3: GEV($\mu_t=\mu$, $\sigma_t=\exp(\sigma_0+\alpha_t)$, ξ)
- Model 4: GEV($\mu_t=\mu_0+\beta_t$, $\sigma_t=\exp(\sigma_0+\alpha_t)$, ξ)
- Model 5: GEV($\mu=\mu_0+\beta_1 \times$ (Nino3.4) , σ, ξ)
- Model 6: GEV(μ, $\log(\sigma)=\sigma_0+\alpha_1 \times$ (Nino3.4), ξ)
- Model 7:GEV($\mu=\mu_0+\beta_1 \times$ (SOI) , σ, ξ)
- Model 8: GEV(μ, $\log(\sigma)=\sigma 0+\alpha 1 \times$ (SOI), ξ)

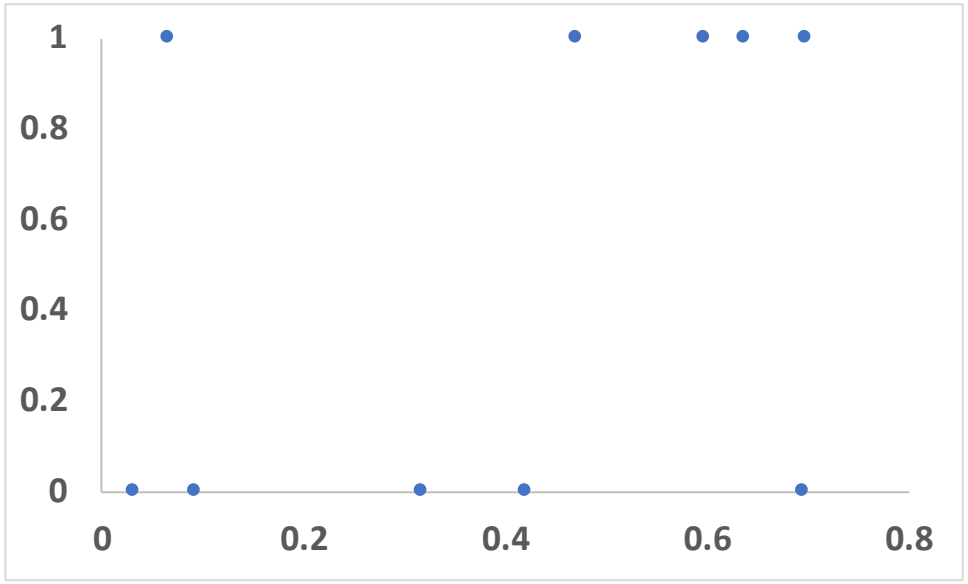
RX1day	AIC	BIC
Model 1	463.3476	468.7676
Model 2	465.343	472.5697
Model 3	465.3544	472.581
Model 4	465.3311	474.3644
Model 5	464.2422	471.4689
Model 6	464.6831	471.9097
Model 7	464.8251	472.0517
Model 8	464.9575	472.1842

R95p	AIC	BIC
Model 1	647.1088	652.5288
Model 2	649.1072	656.3338
Model 3	649.113	656.3397
Model 4	651.1107	660.144
Model 5	647.1196	654.3463
Model 6	648.6866	655.9133
Model 7	649.0021	656.2287
Model 8	648.8509	656.0776

Point-Biserial Correlation

Can we still calculate Correlation coefficient for this case??

Forecast Probability	Rain
0.4229	No
0.0942	No
0.5985	Yes
0.4709	Yes
0.6959	No
0.6999	Yes
0.6385	Yes
0.0336	No
0.0688	Yes
0.3196	No



NOTES

CORRELATION BETWEEN A DISCRETE AND A CONTINUOUS VARIABLE. POINT-BISERIAL CORRELATION

By Robert F. Tate

University of Washington[1]

1. Introduction and Summary. A problem of some importance in statistical applications, especially in the field of psychology, is that of finding a measure of association between a discrete random variable X, which takes the values 0 and 1, and a continuous random variable Y. The ordinary product-moment correlation coefficient $\rho(X, Y)$ is used for this purpose. It has received the name point-biserial correlation coefficient because of its relation to the biserial correlation coefficient proposed by Karl Pearson for a similar problem. The usual estimator r, based on a random sample (X_i, Y_i), $i = 1, 2, \cdots, n$, is referred to as the sample point-biserial correlation coefficient.

The psychological value of ρ (and hence of r) is that it affords a measure of the degree of association between a trait and a measurable characteristic, usually an ability of some kind. For the ith individual in a random sample of n individuals, X_i has the value 1 if the trait is possessed and Y_i is a measure of the ability in question.

We shall give in Section 2 the appropriate mathematical model, based on normal theory, and the asymptotic distribution of r (Theorem 1), the derivation of which is an elementary application of a well known theorem of Cramér. An important special case of this distribution will be discussed in Section 3, namely that in which X takes the values 0 and 1 with equal probabilities. In this connection a variance-stabilizing transformation will be given (Theorem 2). Numerical work based on this transformation may be carried out with the use of existing tables. In particular, the calculation of confidence limits for ρ is immediate. Theorem 2 is especially useful in investigating the association between sex and some other characteristic, since animal populations consist of approximately half males and half females. As an illustration of the ease with which calculations may be carried out, a problem is considered in which the trait is male and the characteristic is IQ.

The small-sample distribution of r is quite easily found, although it is difficult to deal with when n is even moderately large, asymptotic methods appearing to be more desirable. This is discussed in Section 4.

Received 6/29/53, revised 1/26/54.

[1] This research was performed while the author was at the Statistical Laboratory, University of California, Berkeley, and was sponsored in part by the Office of Naval

Acharya, et al (in prep.)

Point biserial correlation

X and Y, where Y is in interval or ratio scale with normal distribution while X is a naturally dichotomous variable.

Forecast Probability	Rain
0.4229	No
0.0942	No
0.5985	Yes
0.4709	Yes
0.6959	No
0.6999	Yes
0.6385	Yes
0.0336	No
0.0688	Yes
0.3196	No

$$r_{pb} = \frac{(\bar{Y}_1 - \bar{Y}_0)}{Sd_Y}\sqrt{\frac{p_X}{q_X}}$$

$\bar{Y}_0$ = mean of Y for individuals scoring o on X.
$\bar{Y}_1$ = mean of Y for individuals scoring 1 on X.

Sdy = the standard deviation of the continuous data.
px = proportion of samples in group 0.
qx = proportion of samples in group 1.

Significance Testing

$$t = \sqrt{N-2}\frac{r_{pb}}{\sqrt{1-r_{pb}^2}}$$

If X is a artificially occurring dichotomous variable

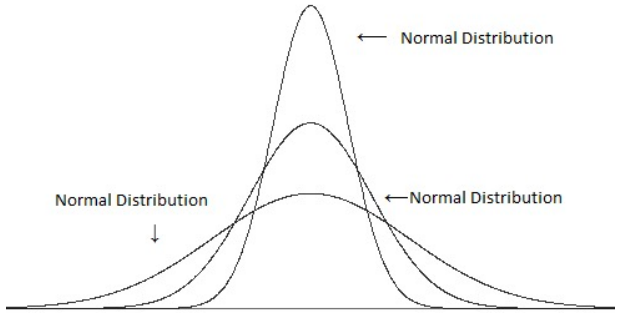
Forecast	Rain<33%le
0.4229	No
0.0942	No
0.5985	Yes
0.4709	Yes
0.6959	No
0.6999	Yes
0.6385	Yes
0.0336	No
0.0688	Yes
0.3196	No

Biserial correlation

The biserial correlation coefficient can also be computed from the point-biserial correlation coefficient

$$r_b = r_{pb}\frac{\sqrt{p_X\, q_X}}{(\lambda)}$$

λ is the ordinate (height) of standardized normal distribution



Numerically, r_b obtained is always greater than r_{pb}.

Is it a very “new” in climate field?

Geofísica Internacional (2002), Vol. 41, Num. 2, pp. 203-212

Biserial correlation between vorticity field and precipitation: Rainfall diagnosis and prediction

Nora E. Ruiz
Depto. de Ciencias de la Atmósfera y los Océanos, Fac. de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina

Received: January 29, 1999; accepted: January 22, 2002.

RESUMEN

Este trabajo concierne al examen de una metodología de la climatología sinóptica, la técnica de correlación biserial, que permite investigar, en este caso, la interrelación entre la circulación atmosférica y la precipitación. Se analiza el significado de los campos de correlación biserial obtenidos relacionando distintas variables representativas del flujo de escala sinóptica, particularmente campos de vorticidad, con la precipitación local, con el propósito de ahondar en metodologías que sean simples, eficientes y fáciles de interpretar para ligar la circulación de gran escala con la pequeña escala o local. Se propone una interpretación basada en las configuraciones de los campos de correlación biserial entre vorticidad en 500 hPa y precipitación, que tiene en cuenta los gradientes de vorticidad anómala con los efectos de cortante y curvatura involucrados, para elucidar posibles mecanismos que favorecen la ocurrencia de lluvia. Las anomalías en la curvatura de los sistemas sinópticos son en gran medida responsables de la precipitación. Se utiliza como ejemplo la precipitación diaria de Córdoba, Argentina, para ilustrar los resultados. Se puede identificar claramente la posición de los centros anómalos de vorticidad ciclónica y anticiclónica y de la corriente en chorro en asociación con la ocurrencia de precipitación. El análisis se hace extensivo para precipitaciones más copiosas.

PALABRAS CLAVE: Precipitación, vorticidad, correlación biserial.

ABSTRACT

This work concerns the examination of a methodology of synoptic climatology, the biserial correlation technique, which allows studying the relationship between atmospheric circulation and precipitation. The physical meaning of biserial correlation fields between variables representing synoptic-scale circulation, particularly vorticity fields, and local precipitation is explored. One purpose is to examine this approach used to link the large-scale circulation and the smaller-scale surface environment, which seems to be simple, efficient and easy to interpret. An analysis based on biserial correlation configurations between 500 hPa vorticity and precipitation takes into account anomalous vorticity gradients including curvature and shear effects to describe some mechanisms favoring the occurrence of rainfall. It is shown that anomalies in the curvature of synoptic systems are largely causing precipitation. Daily precipitation at Córdoba, Argentina is used as an example to illustrate the results. The position of the cyclonic and anticyclonic anomaly centers and the position of the jet streams in association with precipitation may be clearly identified. The analysis is made extensive to heavier rainfall.

KEY WORDS: Precipitation, vorticity, biserial correlation.

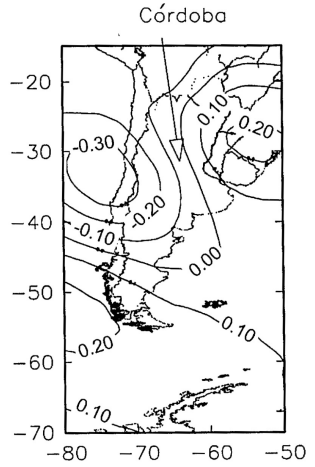


Fig. 1. Biserial correlation field between 500 hPa geopotential heights and the daily precipitation ≥0.1 mm at Córdoba (see text), during the austral summer (November to April).

Extreme floods in central Europe over the past 500 years: Role of cyclone pathway “Zugstrasse Vb”

M. Mudelsee[1]
Department of Earth Sciences, Boston University, Boston, Massachusetts, USA

M. Börngen and G. Tetzlaff
Institute of Meteorology, University of Leipzig, Leipzig, Germany

U. Grünewald
Institute of Hydrology, Technical University Cottbus, Cottbus, Germany

Received 18 May 2004; revised 24 August 2004; accepted 16 September 2004; published 2 December 2004.

[1] Anthropogenically induced climate change has been hypothesized to add to the risk of extreme river floods because a warmer atmosphere can carry more water. In the case of the central European rivers Elbe and Oder, another possibility that has been considered is a more frequent occurrence of a weather situation of the type “Zugstrasse Vb,” where a low-pressure system travels from the Adriatic region northeastward, carrying moist air and bringing orographic rainfall in the mountainous catchment areas (Erzgebirge, Sudeten, and Beskids). Analysis of long, homogeneous records of past floods allows us to test such ideas. M. Mudelsee and co-workers recently presented flood records for the middle parts of the Elbe and Oder, which go continuously back to A.D. 1021 and A.D. 1269, respectively. Here we review the reconstruction and assess the data quality of the records, which are based on combining documentary data from the interval up to 1850 and measurements thereafter, finding both the Elbe and Oder records to provide reliable

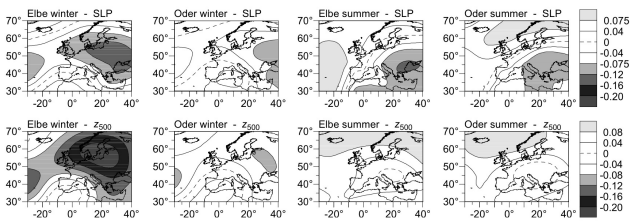


Figure 12. Contour maps of the point-wise biserial correlation coefficient between flood events (Elbe, Oder; winter, summer; classes 1−3) on the one hand and sea level pressure (SLP) or 500 hPa geopotential height (z_{500}) time series on the other; time interval, 1658−1999. Significant correlations (section 3.3) are on color scale. A negative (positive) correlation indicates a pressure below (above) the seasonal average at a geographic point during floods. Elbe and Oder catchment areas are located around 50°N, 15°E (Figure 2). See color version of this figure at back of this issue.

Exploring the Predictability of 30-Day Extreme Precipitation Occurrence Using a Global SST–SLP Correlation Network

MENGQIAN LU AND UPMANU LALL
Department of Earth and Environmental Engineering, and Columbia Water Center, Columbia University, New York, New York

JAYA KAWALE
Department of Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota

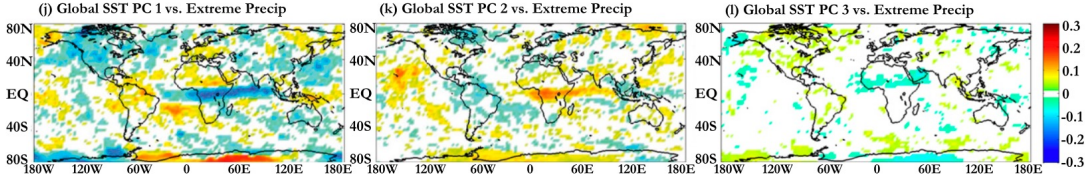
STEFAN LIESS
Department of Soil, Water, and Climate, University of Minnesota, St. Paul, Minnesota

VIPIN KUMAR
Department of Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota

(Manuscript received 16 May 2014, in final form 2 November 2015)

ABSTRACT

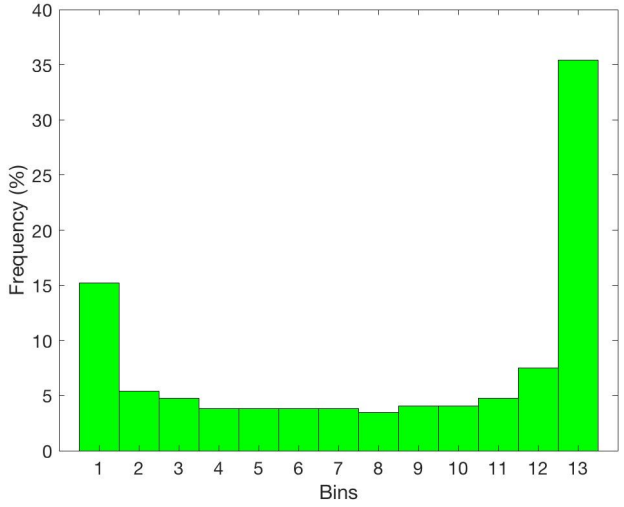
Correlation networks identified from financial, genomic, ecological, epidemiological, social, and climatic data are being used to provide useful topological insights into the structure of high-dimensional data. Strong convection over the oceans and the atmospheric moisture transport and flow convergence indicated by atmospheric pressure fields may determine where and when extreme precipitation occurs. Here, the spatiotemporal relationship among sea surface temperature (SST), sea level pressure (SLP), and extreme global precipitation is explored using a

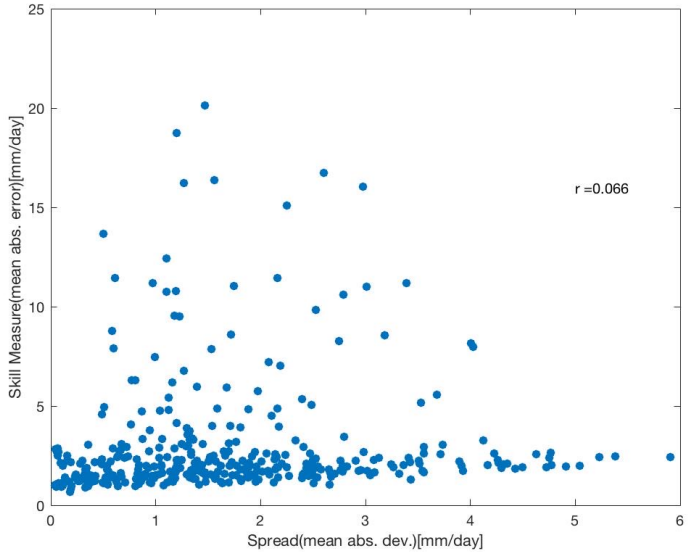


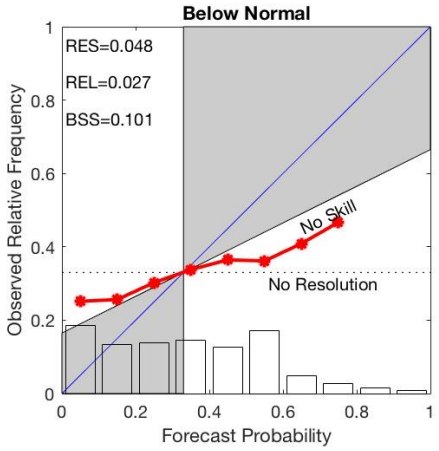
Mostly used for “teleconnection” study but not for forecast verification

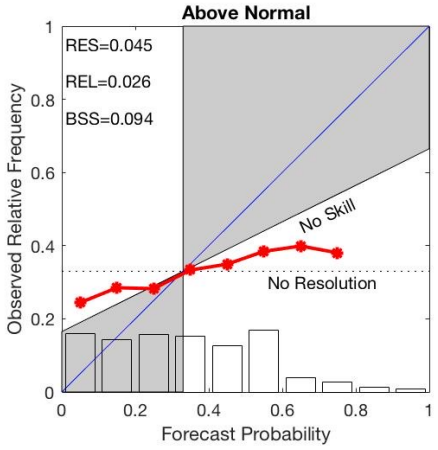
Example: Measuring Probabilistic Seasonal forecast for Indian Monsoon

Variable: Precipitation
Season: Jun-Jul-Aug-Sep
forecast at: May
Period:1982 to 2010
GCM: GFDL-CM2p5-FLOR-B01
Method: Counting member

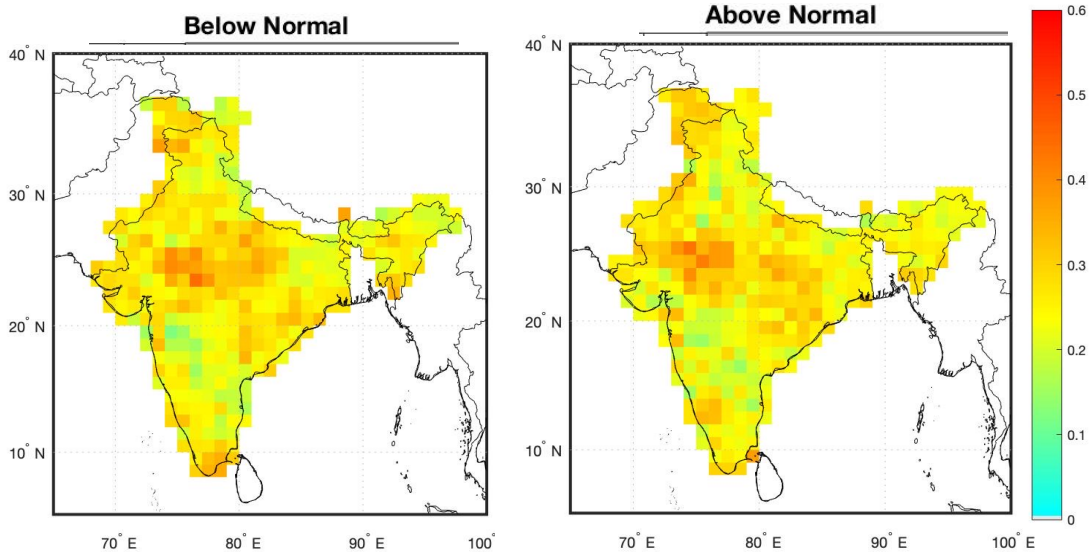






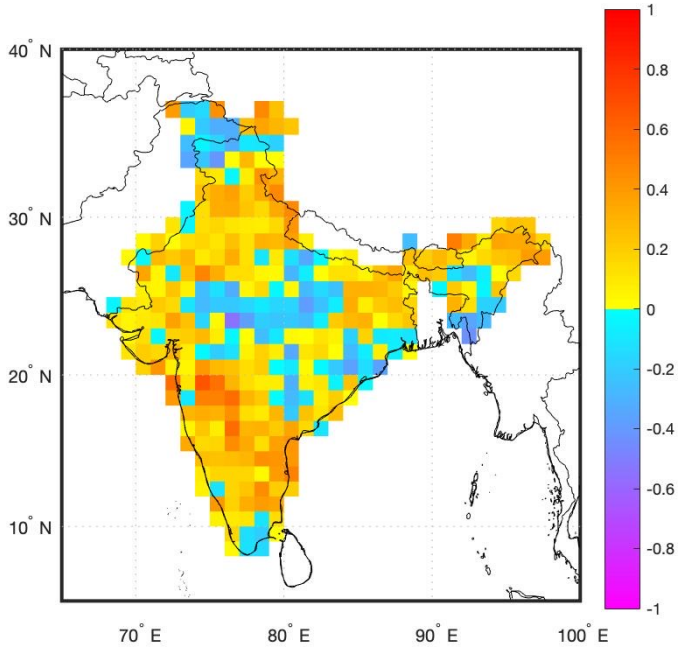


Brier Score

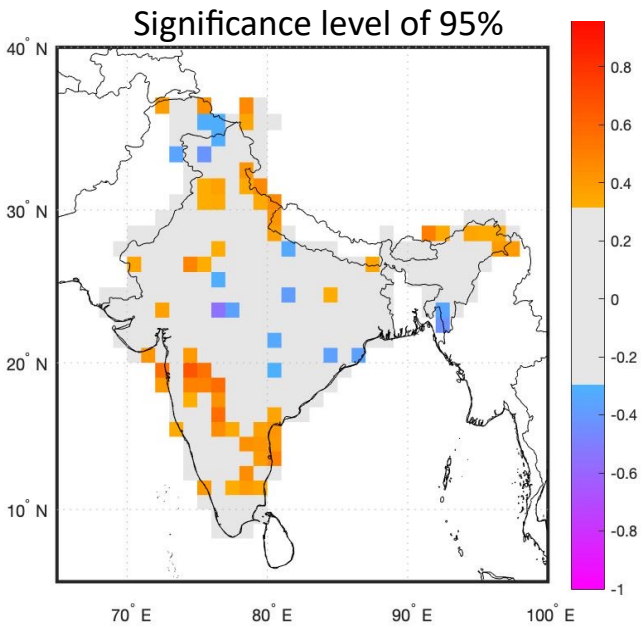


Point Biserial correlation

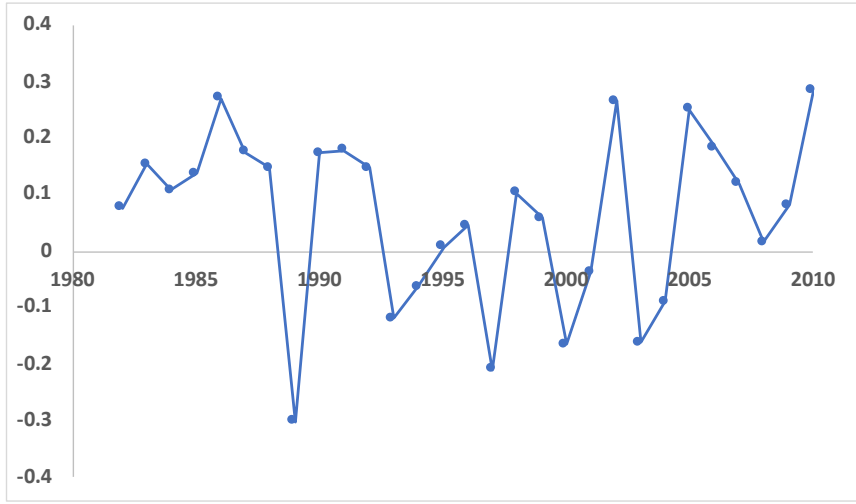
Below Normal



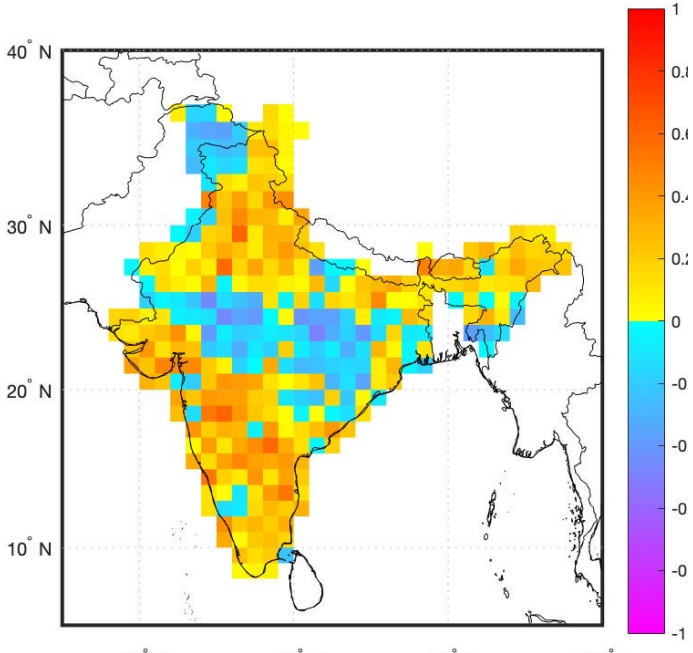
Significance level of 95%

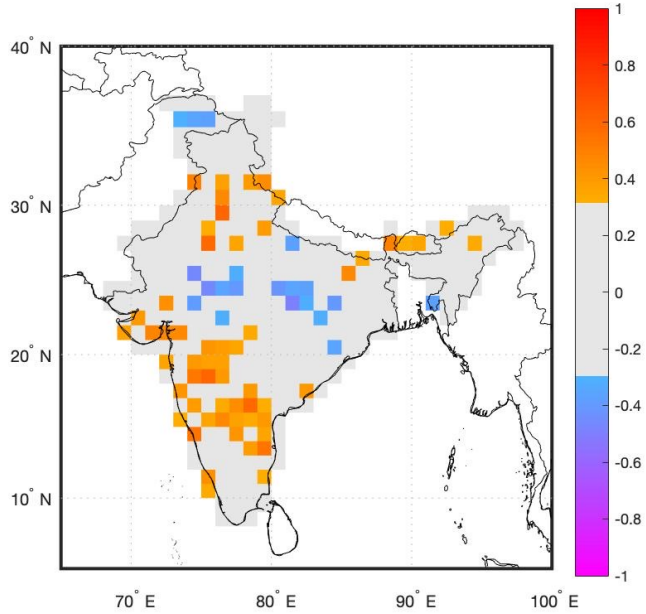


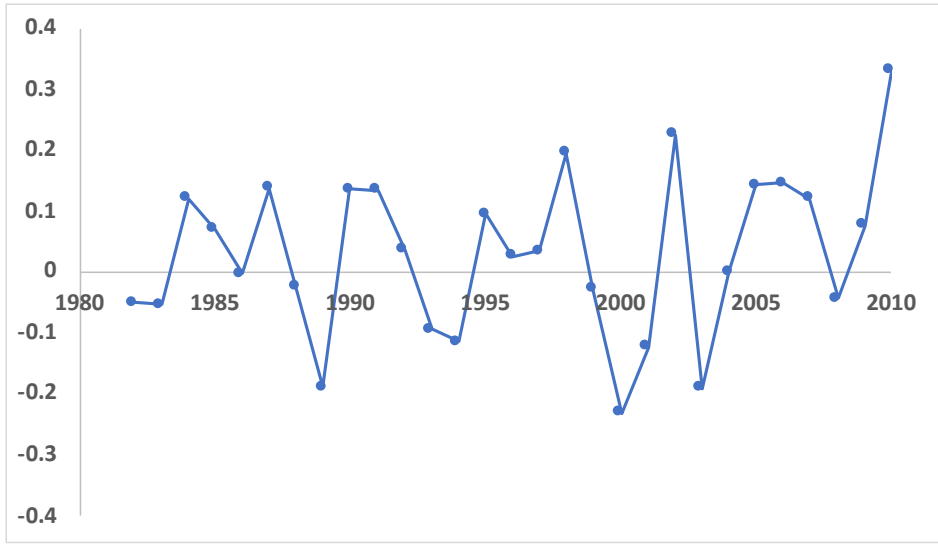
Area Average (~ACC)



Above Normal







Future Direction
Relationship with Brier Score/ Brier Skill Score

VOLUME 116 MONTHLY WEATHER REVIEW DECEMBER 1988

Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient

ALLAN H. MURPHY

Department of Atmospheric Sciences, Oregon State University, Corvallis, Oregon

(Manuscript received 1 February 1988, in final form 11 April 1988)

ABSTRACT

Several skill scores are defined, based on the mean-square-error measure of accuracy and alternative climatological standards of reference. Decompositions of these skill scores are formulated, each of which is shown to possess terms involving 1) the coefficient of correlation between the forecasts and observations, 2) a measure of the nonsystematic (i.e., conditional) bias in the forecasts, and 3) a measure of the systematic (i.e., unconditional) bias in the forecasts. Depending on the choice of standard of reference, a particular decomposition may also contain terms relating to the degree of association between the reference forecasts and the observations. These decompositions yield analytical relationships between the respective skill scores and the correlation coefficient, document fundamental deficiencies in the correlation coefficient as a measure of performance, and provide additional insight into basic characteristics of forecasting performance. Samples of operational precipitation probability and minimum temperature forecasts are used to investigate the typical magnitudes of the terms in the decompositions. Some implications of the results for the practice of forecast verification are discussed.

1. Introduction

Skill scores are generally defined as measures of the relative accuracy of forecasts produced by two forecasting systems, one of which is a "reference system" (e.g., see Murphy and Daan 1985). Positive skill (i.e., a favorable difference in accuracy) is usually considered to represent a minimal level of acceptable performance for a set of forecasts. To the extent that the difficulty inherent in forecasting situations is reflected in the level of accuracy of the reference forecasts, skill scores also take difficulty into account. As a result, they can be used (with appropriate caveats) to compare forecasting performance across different locations or time periods. Thus, it is not surprising that skill scores are widely used in evaluating the performance of operational and experimental forecasts (e.g., see Dagostaro et al. 1988; Murphy and Daan 1985).

In the context of forecast verification, correlation coefficients are measures of the degree of linear association in conjunction with model verification studies (e.g., see Arpe et al. 1985; Miyakoda et al. 1972; Sanders 1987).

Despite the rather widespread use of both skill scores and correlation coefficients, the relationships between these two common types of verification measures have evidently not been explored. In addition, little if any attention has been devoted to the problem of obtaining a quantitative appreciation of the deficiencies in the correlation coefficient as a measure of forecasting performance. The primary purpose of this paper is to describe decompositions of a family of climatological skill scores that yield insight into (i) the relationships between these measures and the (product moment) correlation coefficient and (ii) the deficiencies in the latter as a performance measure.

In section 2, we define the terms "accuracy" and "skill" and identify the basic measures of these attributes—namely, the mean-square-error m
curacy and the mean-square-error skill
ployed in this paper. This section also de

$$\mathrm{SS}(f, \bar{x}, x) = r_{fx}^2 - [r_{fx} - (s_f/s_x)]^2 - [(\bar{f} - \bar{x})/s_x]^2. \quad (12)$$

Forecasts $f_1, \ldots, f_N$.
Observations $o_1, \ldots, o_N$.

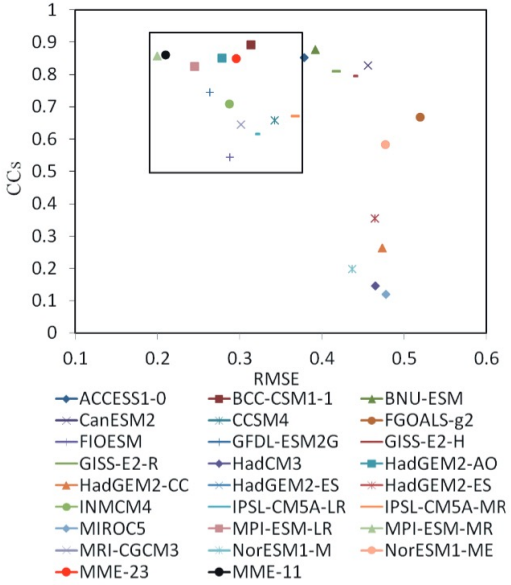
$$\mathrm{MSESS} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}_{\mathrm{ref}}} = 1 - \frac{\sum_{n=1}^{N}(f_n - o_n)^2}{\sum_{n=1}^{N}(o_n - \bar{o})^2} = 1 - \frac{\mathrm{MSE}}{\hat{\sigma}_o^2}$$

For this reference forecast $\bar{o}$, the decomposition of MSESS is

$$\mathrm{MSESS} = \mathrm{AC}^2 - \left(\frac{\hat{\sigma}_f}{\hat{\sigma}_o} - \mathrm{AC}\right)^2 - \frac{(\bar{f} - \bar{o})^2}{\hat{\sigma}_o^2}$$

$$\mathrm{MSESS} = \text{Frac. explained var.} - \text{Cond. bias}^2 - \text{Bias}^2$$

Application: to choose "good models"



Yajuan et al,2015

$$BS = REL - RES + UNC$$

Each of these components can be decomposed further according to the number of p

$$BS = \frac{1}{N}\sum_{k=1}^{K} n_k(\mathbf{f_k} - \bar{\mathbf{o}}_\mathbf{k})^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{\mathbf{o}}_\mathbf{k} - \bar{\mathbf{o}})^2 + \bar{\mathbf{o}}\,(1 - \bar{\mathbf{o}})$$

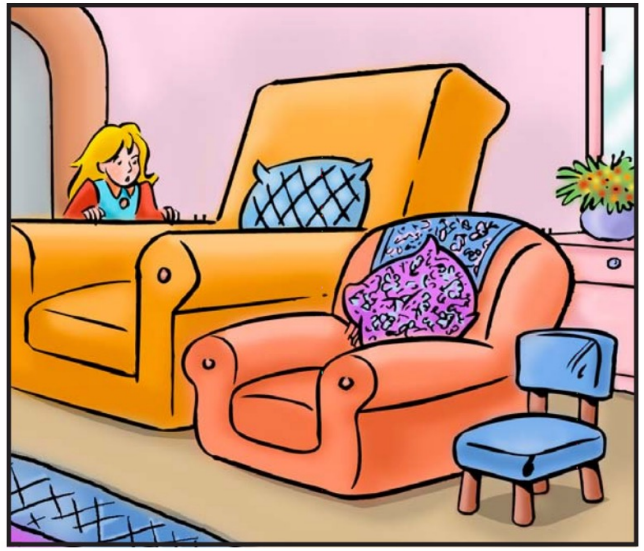
"The models with CCs greater than 0.5 and RMSEs less than 0.37 are selected to pro-duce the "best model ensemble"…"

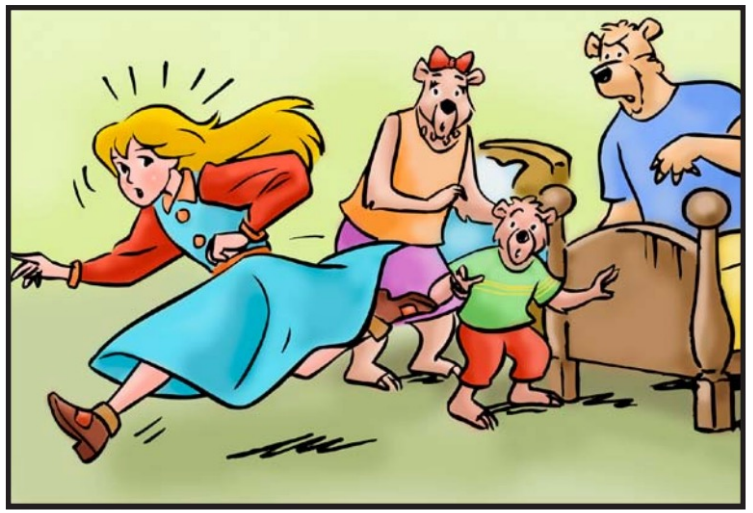
References:

- Short Communication A refined index of model performance(https://climate.sitehost.iu.edu/RobesonPubs/willmott_etal_ijoc_2012.pdf)
- Roulston, M. S., & Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. Monthly Weather Review, 130, 1653–1660.
- Forecast Comparison Based on Random Walks (http://cola.gmu.edu/delsole/webpage/SkillComparison/skillcomparison.html)
- Interpreting the skill score form of forecast performance metrics (https://www.sciencedirect.com/science/article/abs/pii/S0169207019300093)

Let’s have some story time!!

Goldilocks and the Three Bears





Thanks!!
Any Question??
npa5302@psu.edu