Polychoric correlation coefficient in forecast verification based on KxK contingency tables

***Zoran Pasarić* and Josip Juras**

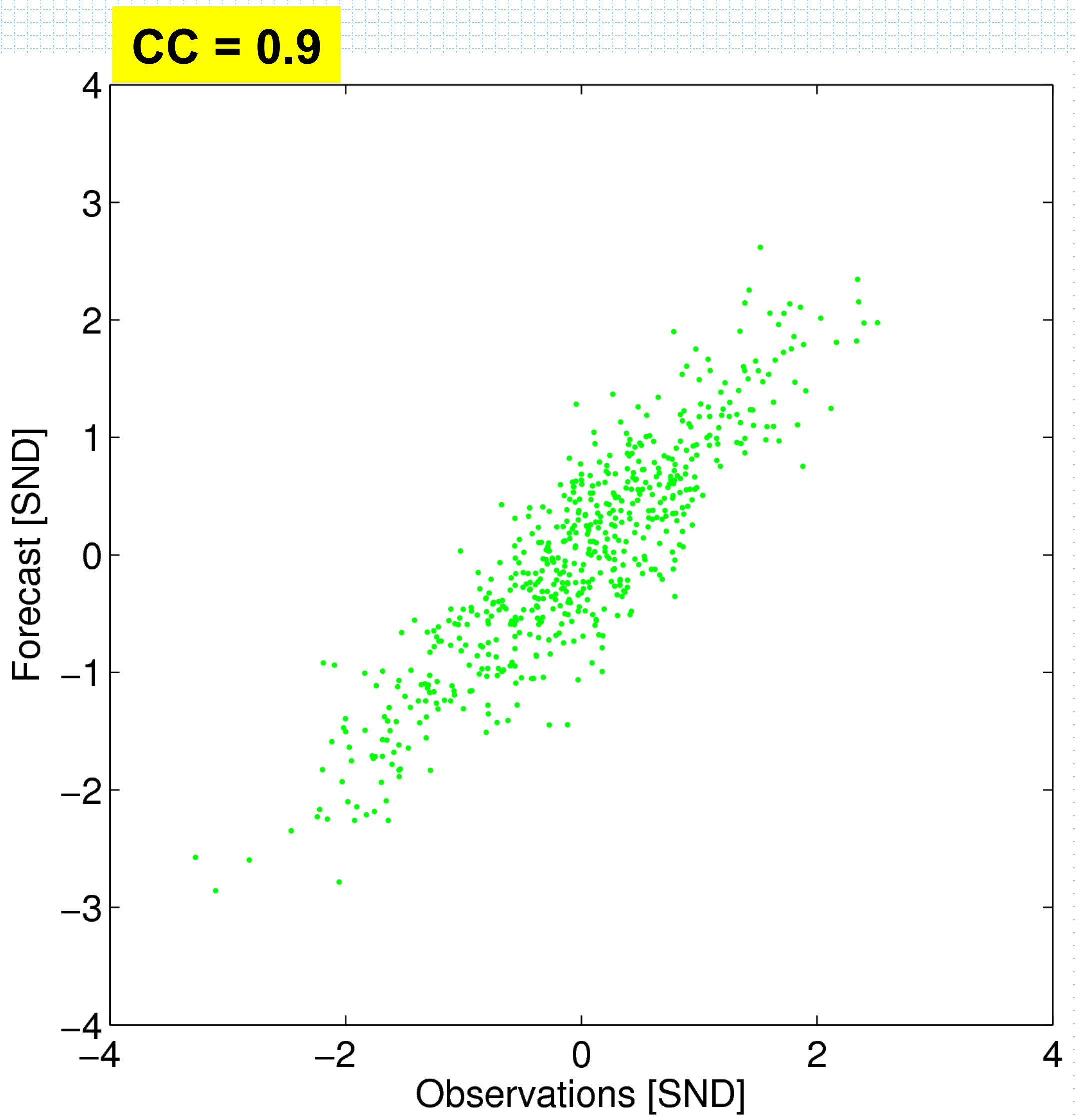
Geophysical Institute, Faculty of Science
University of Zagreb, CROATIA

Fourth International Verification Methods Workshop
Helsinki, 8 -10 June 2009

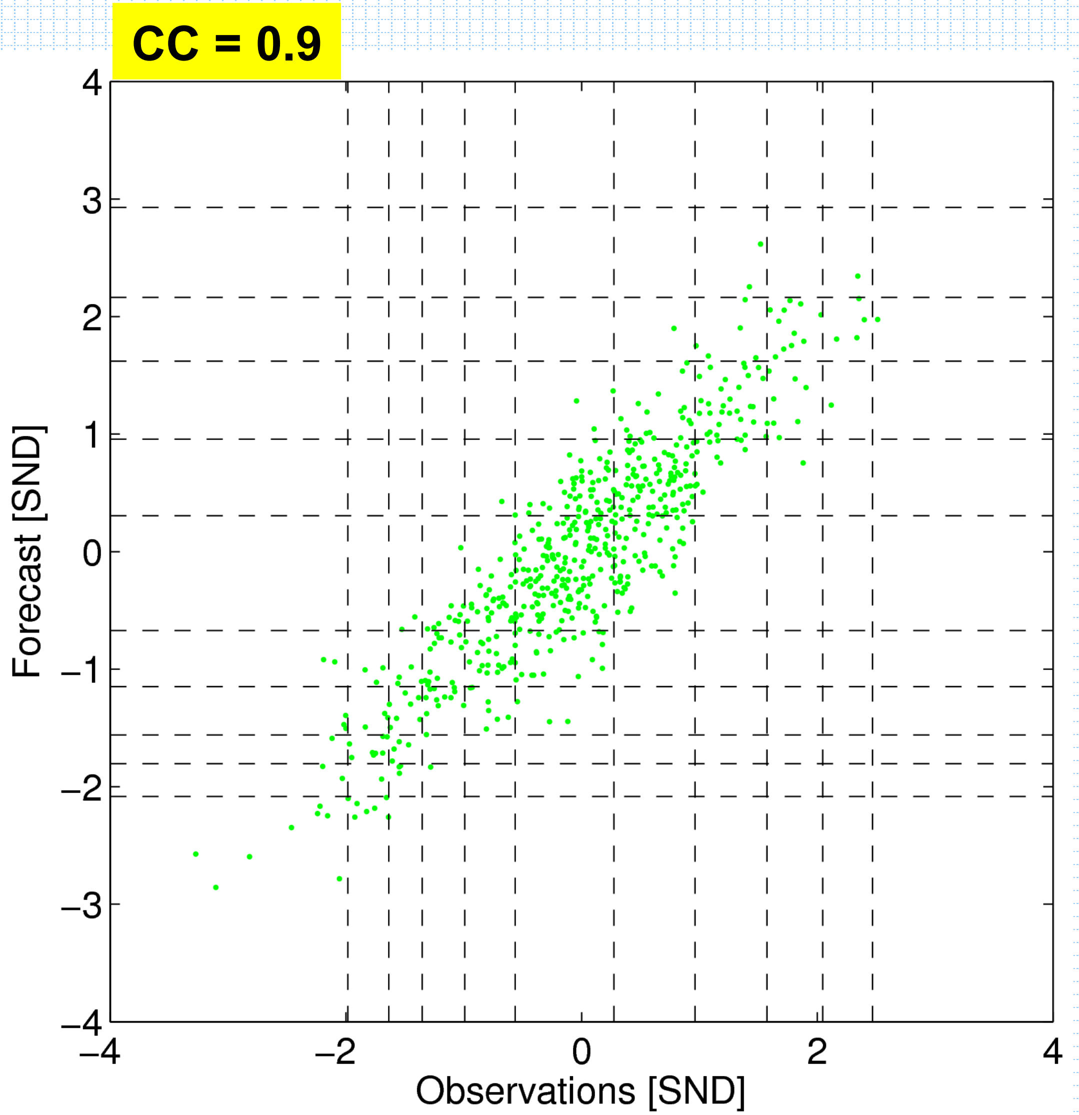
Outline

- The bivariate normal distribution (BND) and KxK table
- Example:
  - PCC for 11x11 tables of temperature change forecasts
  - Additional information: Biases and base rates
  - Reconstruction
  - Residual
- Summary of PCC
- More examples
  - QPF for the United States (6x6 tables)

Bivariate normal distribution (BND) and K x K contingency table



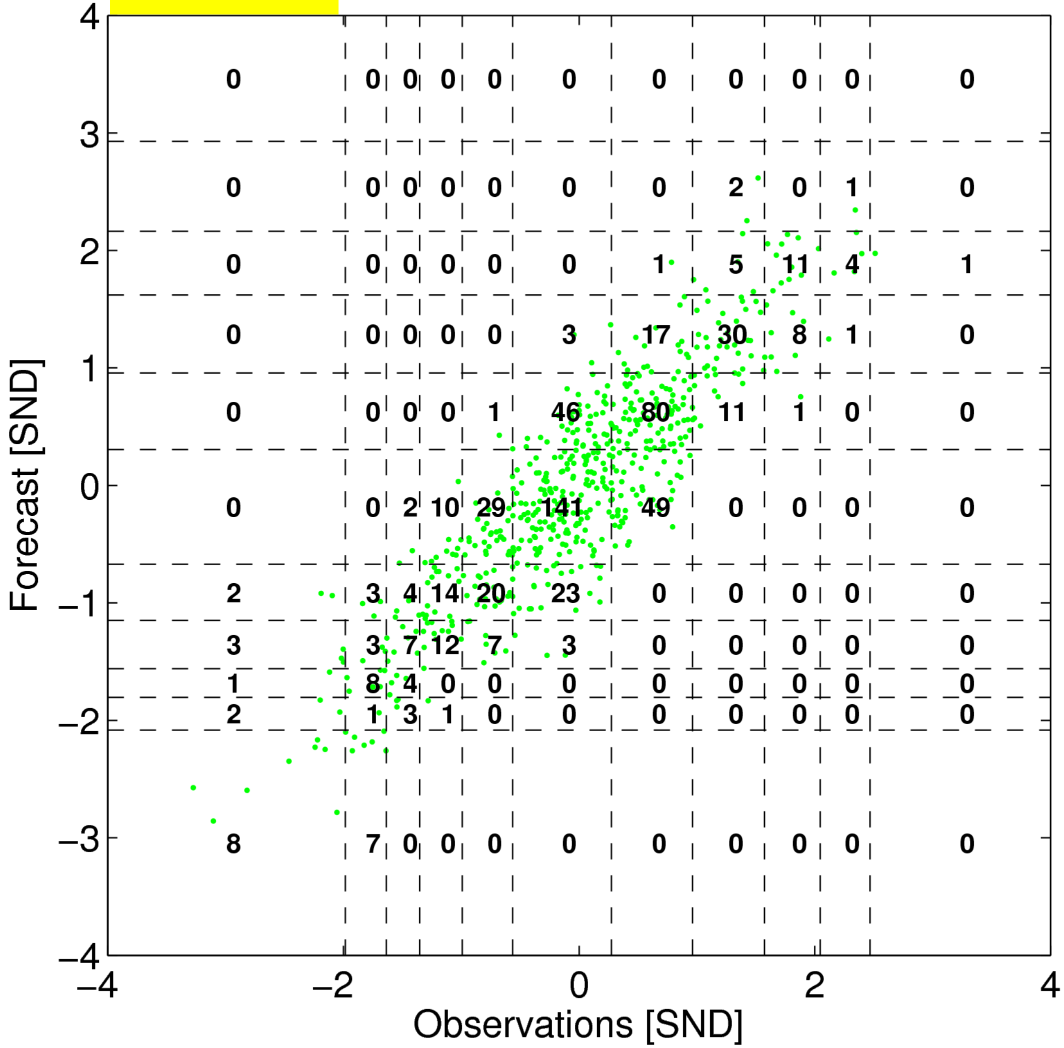
Bivariate normal distribution (BND) and K x K contingency table



- From CC towards the table

Bivariate normal distribution (BND) and K x K contingency table

CC = 0.9



- From CC towards the table

- From table towards the CC (ML method)

- Polychoric Corelation Coefficient, PCC (*Ritchie-Scott*, 1918, *Pearson*,1922)

Example

Brooks & Doswell (W&F,1996): Four 11 x 11 tables of temperature changes

Observation

Forecast	COLDER					N.C.					WARMER
COLDER	8	3	0	0	0	0	0	0	0	0	0
	4	6	0	0	0	0	0	0	0	0	0
	1	3	9	1	0	0	0	0	0	0	0
	1	3	10	20	5	0	0	0	0	0	0
	0	1	2	14	39	16	2	0	0	0	0
N.C.	0	0	1	7	24	139	43	4	0	0	0
	0	0	0	1	4	30	65	22	2	0	0
	0	0	0	0	2	5	22	32	6	2	0
WARMER	0	0	0	0	0	0	0	7	12	2	1
	0	0	0	0	0	0	0	0	2	4	2
	0	0	0	0	0	0	0	0	0	0	1

Forecasting system	CC (from B&D)	TCC (ML method)
NWSFO	0.91	0.902

Example

Brooks & Doswell (W&F,1996): Four 11 x 11 tables of temperature changes

Observation

Forecast \ Observation			COLDER			N.C.			WARMER		
	8	3	0	0	0	0	0	0	0	0	0
COLDER	4	6	0	0	0	0	0	0	0	0	0
	1	3	9	1	0	0	0	0	0	0	0
	1	3	10	20	5	0	0	0	0	0	0
	0	1	2	14	39	16	2	0	0	0	0
N.C.	0	0	1	7	24	139	43	4	0	0	0
	0	0	0	1	4	30	65	22	2	0	0
	0	0	0	0	2	5	22	32	6	2	0
	0	0	0	0	0	0	0	7	12	2	1
WARMER	0	0	0	0	0	0	0	0	2	4	2
	0	0	0	0	0	0	0	0	0	0	1

Forecasting system	CC (from B&D)	TCC (ML method)
NWSFO	0.91	0.902
LFM-MOS	0.87	0.856
NGM-MOS	0.88	0.872
CON	0.90	0.896

Forecasting system	TCC, 5x5 (ML method)
NWSFO	0.903
LFM-MOS	0.848
NGM-MOS	0.856
CON	0.884

Differences: NWSFO - CON

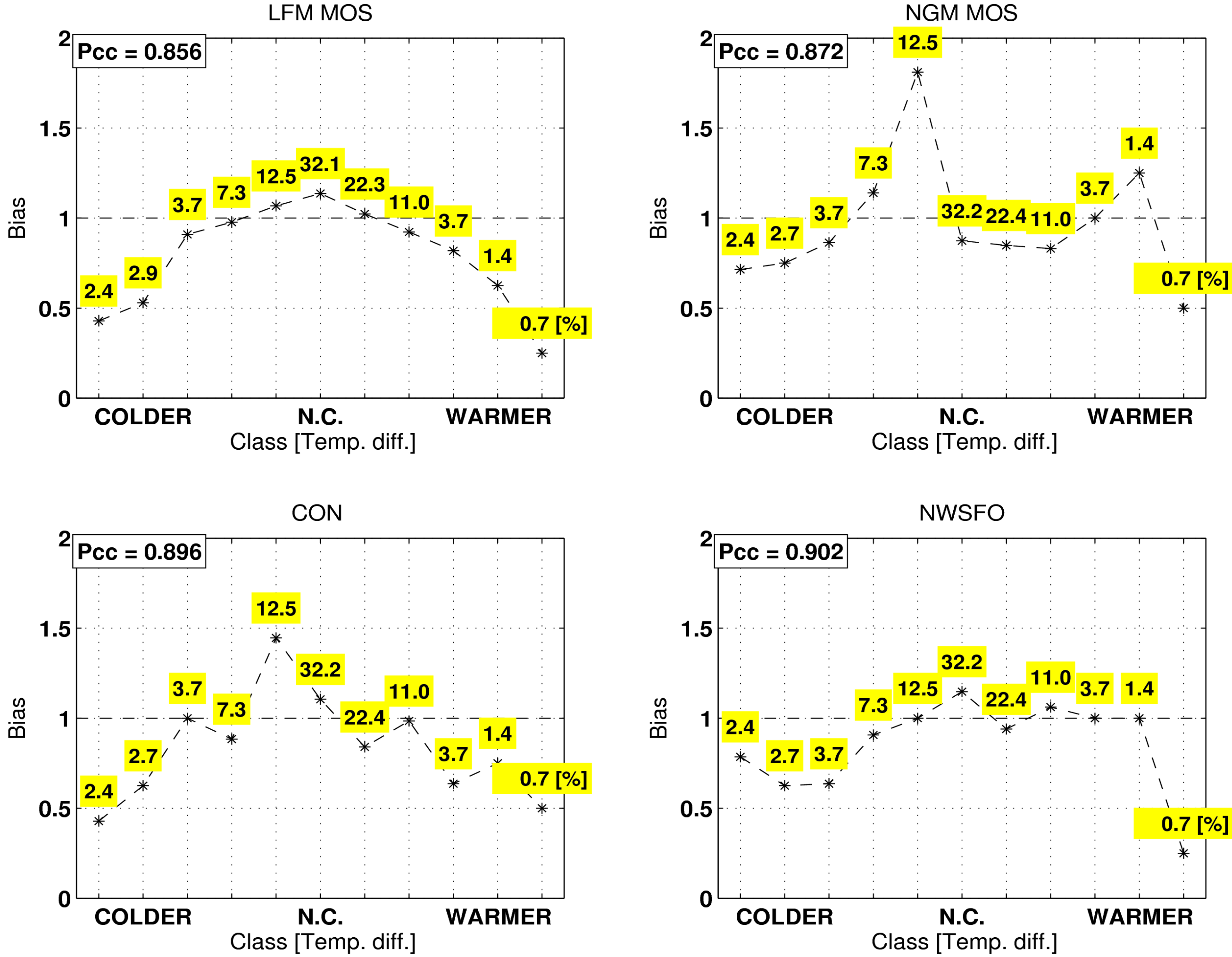
		COLDER				N.C.			WARMER		
	4	1	0	0	0	0	0	0	0	0	0
COLDER	−2	2	0	0	0	0	0	0	0	0	0
	−2	−3	−1	−2	0	0	0	0	0	0	0
	0	1	2	0	0	−2	0	0	0	0	0
	0	0	0	−2	−3	−27	−1	0	0	0	0
N.C.	0	−1	−1	4	1	20	−11	−4	0	0	0
	0	0	0	0	0	6	6	1	0	0	0
	0	0	0	0	2	3	6	−2	−4	0	0
	0	0	0	0	0	0	0	5	3	1	−1
WARMER	0	0	0	0	0	0	0	0	1	0	1
	0	0	0	0	0	0	0	0	0	−1	0

TCC: 0.896 ⟶ 0.902

Additional information:

Biases and marginal frequencies of observations

TCC measures
the association, only



Reconstruction

The KxK contingency table:

Observation

Forecast	C_1	C_2	…	C_K	
C_1	P_{11}	P_{12}	…	P_{1K}	$P_{F,1}$
C_2	P_{21}	P_{22}	…	P_{2K}	$P_{F,2}$
…	…	…	…	…	…
C_K	P_{K1}	P_{K2}	…	P_{KK}	$P_{F,K}$
	$P_{O,1}$	$P_{O,2}$	…	$P_{O,K}$	

- Consider the table obtained by *partitioning* a normalized BND according to some *thresholds*

- From CC and marginal frequencies it is possible to reconstruct the whole table**!**

$$\text{Bias} = (P_{F,1}/P_{1.}, \ \dots, \ P_{F,K-1}/P_{O,K-1}),$$

$$P_O = (P_{O,1}, \ \dots, \ P_{O,K-1})$$

K x K table $\Leftrightarrow$ (TCC, Bias, P_{OBS}) + residual

$K^2 \quad \Rightarrow \quad 1 + (K-1) + (K-1) + 1$

Total no. of elements

The residuals: Overall

Residual table = Original minus theoretical (BND) table

Sums of absolute differences [%]

	LFM MOS	NGM MOS	CON	NWSFO
11 x 11	20.3	20.2	17.4	21.2
5 x 5	15.0	13.5	10.1	14.2
3 x 3	8.6	5.5	4.5	10.8

The residuals, cont.

CON: TCC=0.896, resid=17.4%, N=590

		COLDER				N.C.				WARMER	
	−0.7	1.1	−0.3	−0.1	−0	−0	−0	−0	−0	0	0
COLDER	1.9	1.1	−1.7	−1	−0.2	−0	−0	−0	−0	0	0
	−0.4	0.6	4.6	−2.1	−2.3	−0.4	−0	−0	−0	−0	0
	−0.4	−2.4	0.7	8.3	−4.7	−1.4	−0	−0	−0	−0	−0
	−0.4	−1.3	−4.6	−4	5.3	4	1	−0	−0	−0	−0
N.C.	−0	0.9	1.3	−2	−1.5	−2.9	1.3	3.1	−0.1	−0	−0
	−0	−0	−0	1	3.5	0.3	−0.1	−4	−0.5	−0.2	−0
	−0	−0	−0	−0	−0	0.5	−1.6	3.8	−1.9	−0.4	−0.3
	−0	−0	−0	−0	−0	−0	−0.6	−2.2	3.5	−1.9	1.1
WARMER	0	0	−0	−0	−0	−0	−0	−0.6	−0.8	1.9	−0.5
	−0	0	0	−0	−0	−0	−0	−0	−0.2	0.5	−0.3

NWSFO: TCC=0.902, resid=21.2%

		COLDER				N.C.				WARMER	
	0.8	0.7	−1	−0.4	−0.1	−0	−0	−0	−0	0	0
COLDER	1.1	3	−2.2	−1.4	−0.4	−0	−0	−0	−0	0	0
	−1	−0.4	5.4	−2.4	−1.4	−0.2	−0	−0	−0	−0	0
	−0.6	−1.9	2	7.7	−4.4	−2.9	−0	−0	−0	−0	−0
	−0.3	−1	−3.9	−2.9	12	−5.3	1.4	−0	−0	−0	−0
N.C.	−0	−0.3	−0.4	−1.5	−10.8	9.1	2.5	1.4	−0	−0	−0
	−0	−0	−0	1	3.1	−3.4	−1.5	0.5	0.4	−0.1	−0
	−0	−0	−0	−0	2	2.7	−1.1	−0.3	−3.7	0.5	−0.1
	−0	−0	−0	−0	−0	−0	−1.3	−0.9	3.7	−1.6	0.2
WARMER	0	0	−0	−0	−0	−0	−0	−0.7	−0.3	1.3	−0.2
	−0	0	0	0	−0	−0	−0	−0	−0	−0.2	0.2

PCC: 0.896 $\longrightarrow$ 0.902

Correction of 2 three-class errors improves the association as correction of 20, or so, one-class errors

Question

- Sampling variability due to insufficient sample size?

or

- Real features of the prognostic system ?

- Measure oriented
- Distribution oriented

Complementary approaches

Summary of PCC

- Partition of information

  K x K table ⇔ (PCC, Bias, P_{OBS}) + residual

- Reduction in dimensionality

  K^2 → 2*K

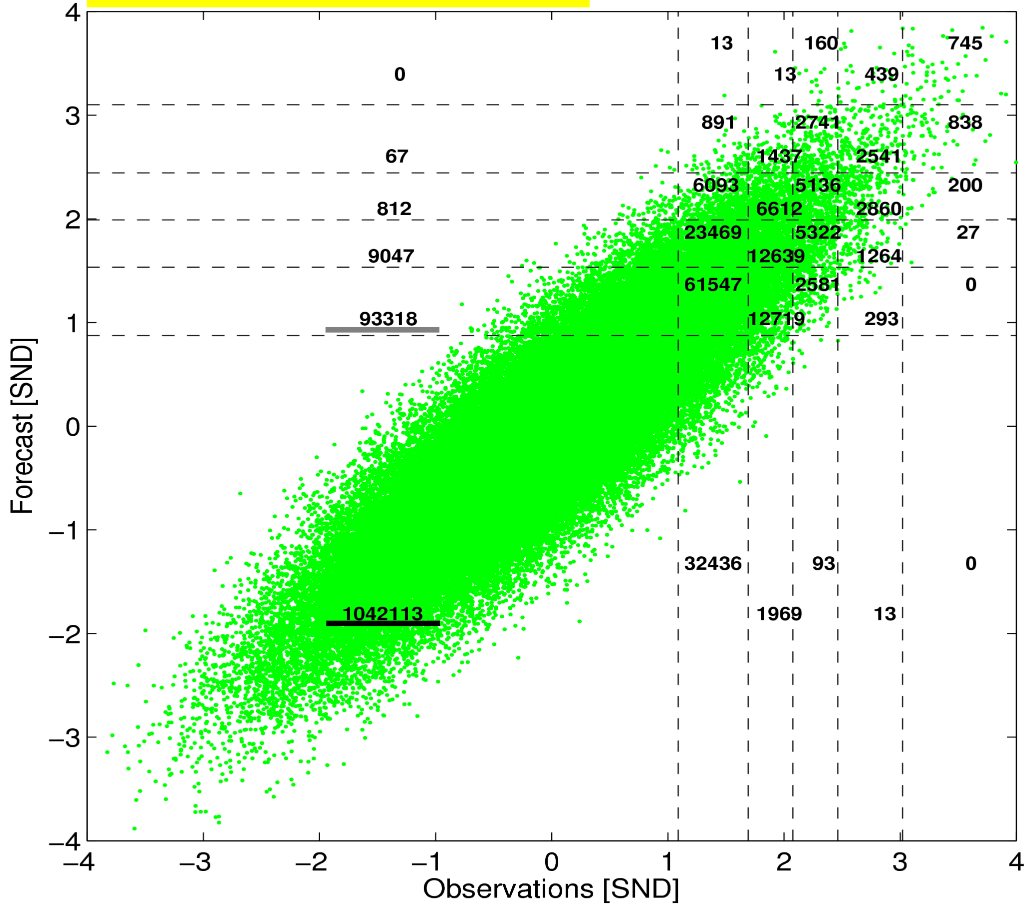
- The PCC, Biases and P_{OBS} are independent of each other

- Using them, the table could be essentially reconstructed

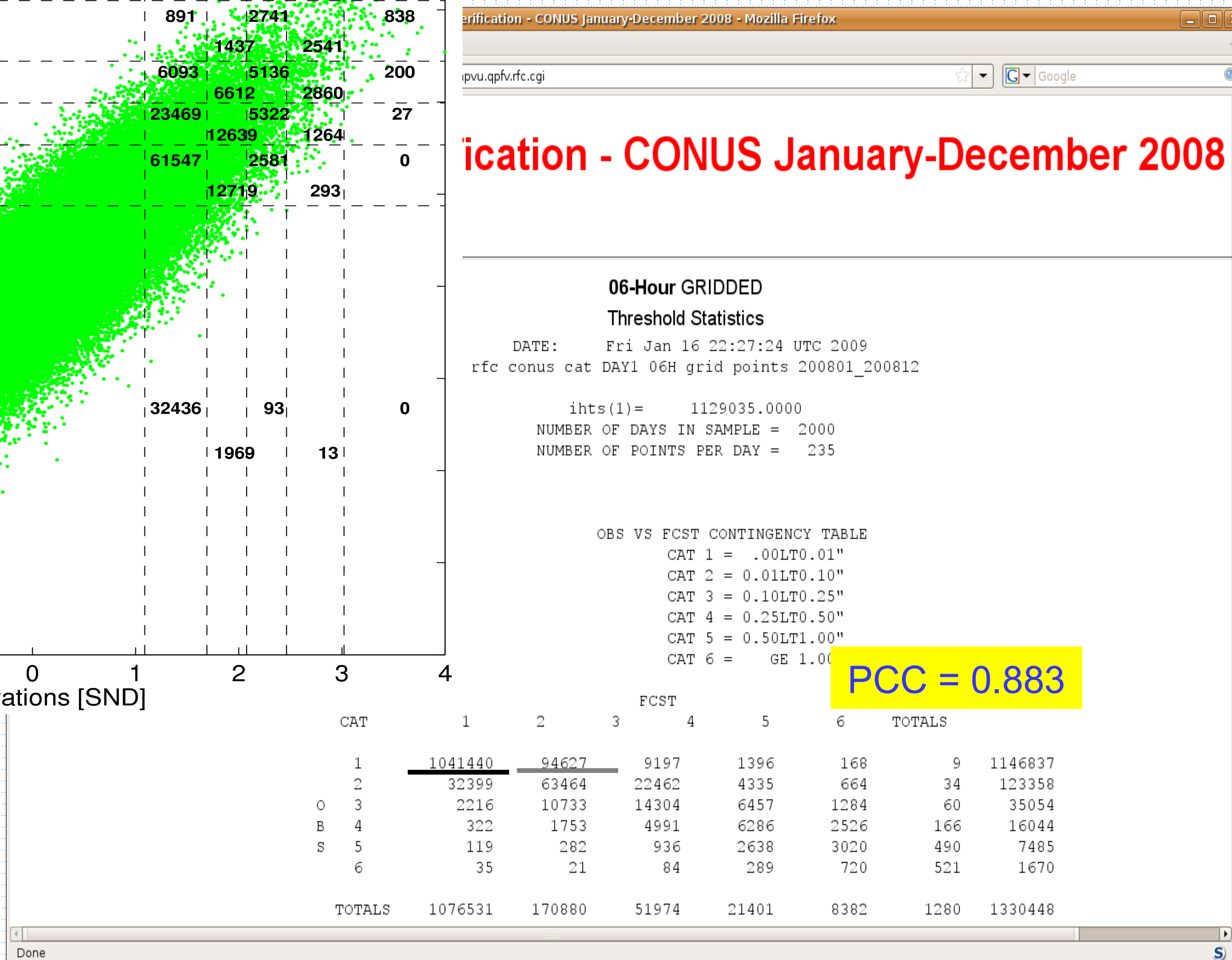
- The distribution oriented approach could be applied to (usually small) residual

More examples
QPF, USA CONUS

Monte Carlo, cc=PCC



http://www.hpc.ncep.noaa.gov/npvu/qpfv/


erification - CONUS January-December 2008 - Mozilla Firefox

npvu.qpfv.rfc.cgi

ication - CONUS January-December 2008

06-Hour GRIDDED

Threshold Statistics

DATE: Fri Jan 16 22:27:24 UTC 2009
rfc conus cat DAY1 06H grid points 200801_200812

ihts(1)= 1129035.0000
NUMBER OF DAYS IN SAMPLE = 2000
NUMBER OF POINTS PER DAY = 235

OBS VS FCST CONTINGENCY TABLE
CAT 1 = .00LT0.01"
CAT 2 = 0.01LT0.10"
CAT 3 = 0.10LT0.25"
CAT 4 = 0.25LT0.50"
CAT 5 = 0.50LT1.00"
CAT 6 = GE 1.00"

FCST

	CAT	1	2	3	4	5	6	TOTALS
	1	1041440	94627	9197	1396	168	9	1146837
	2	32399	63464	22462	4335	664	34	123358
O	3	2216	10733	14304	6457	1284	60	35054
B	4	322	1753	4991	6286	2526	166	16044
S	5	119	282	936	2638	3020	490	7485
	6	35	21	84	289	720	521	1670
	TOTALS	1076531	170880	51974	21401	8382	1280	1330448

Done

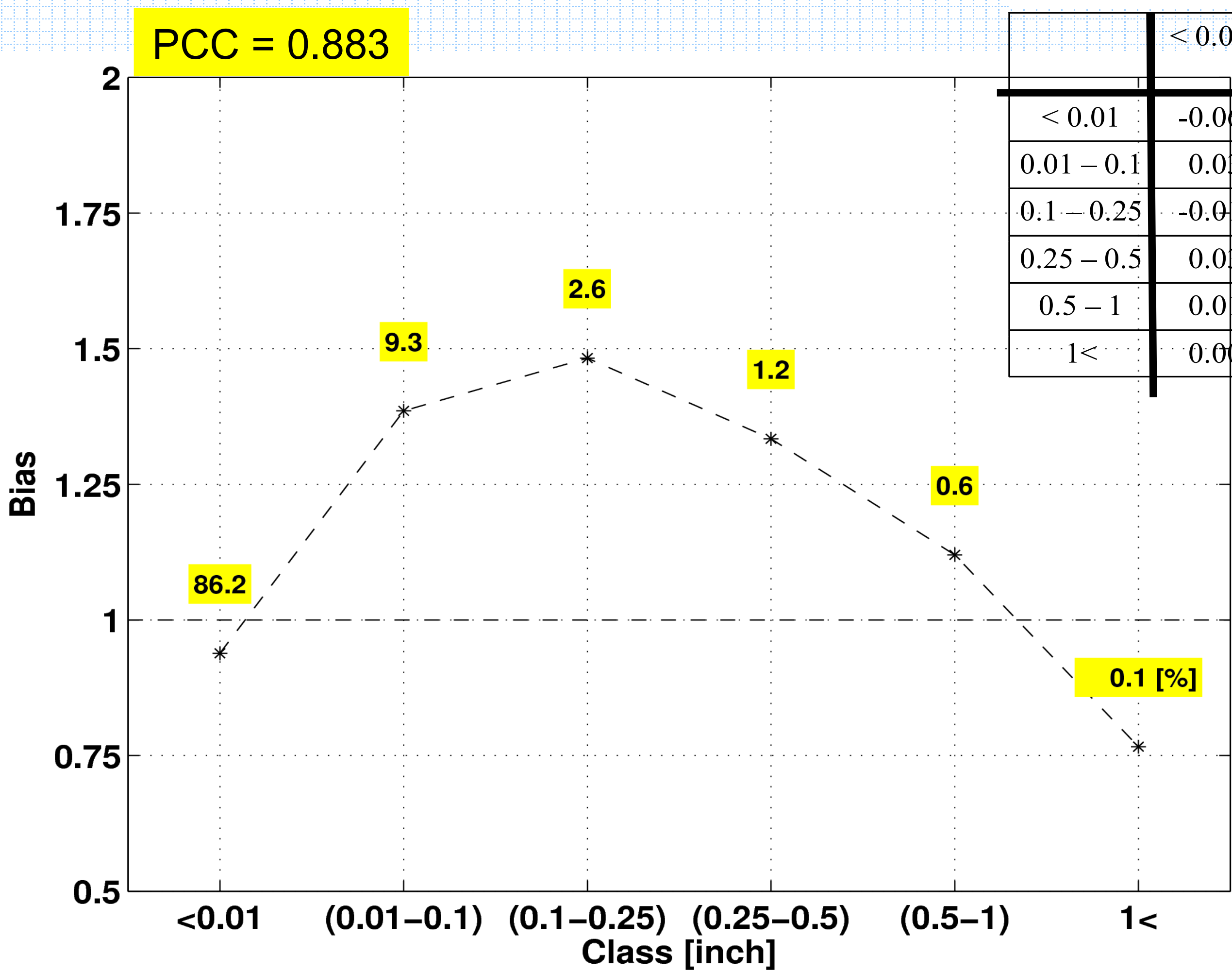
PCC = 0.883

QPF, USA CONUS 2008

Biases, frequencies of observations, residual

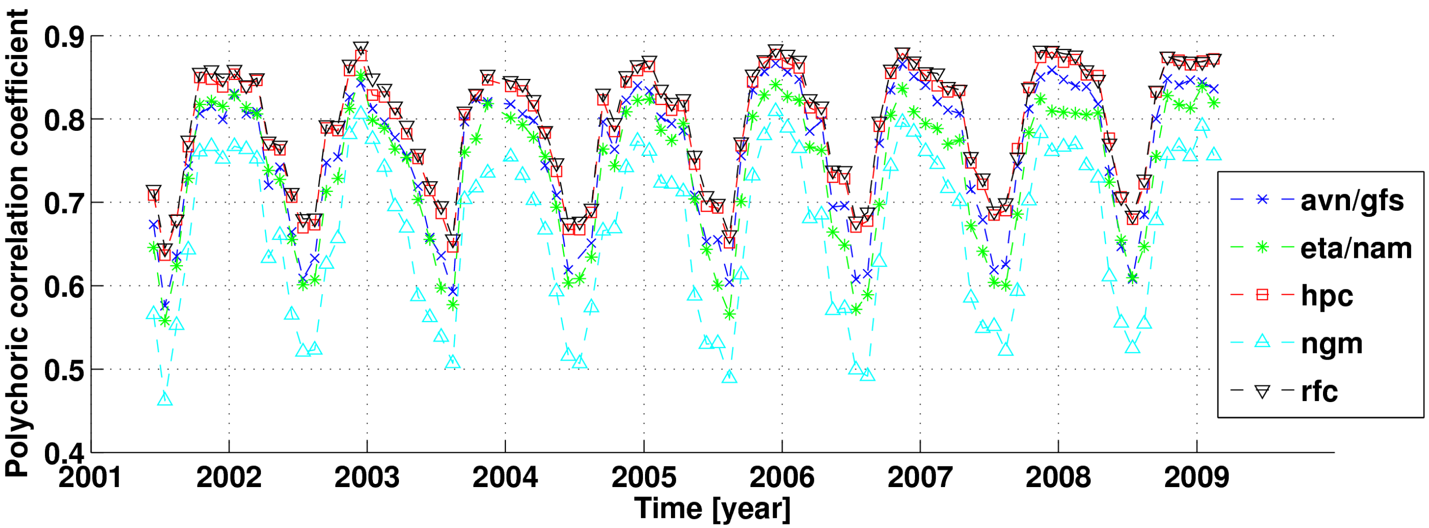
sum(abs(differences)) = 1.15 %

	< 0.01	0.01 – 0.1	0.1 – 0.25	0.25 – 0.5	0.5 – 1	1<
< 0.01	-0.067	0.021	0.024	0.012	0.008	0.003
0.01 – 0.1	0.035	0.167	-0.131	-0.068	-0.003	0.001
0.1 – 0.25	-0.016	-0.074	0.142	-0.024	-0.030	0.002
0.25 – 0.5	0.038	-0.115	-0.010	0.080	0.004	0.003
0.5 – 1	0.010	-0.001	-0.025	0.001	0.020	-0.004
1<	0.001	0.002	0.001	0.001	0.000	-0.005

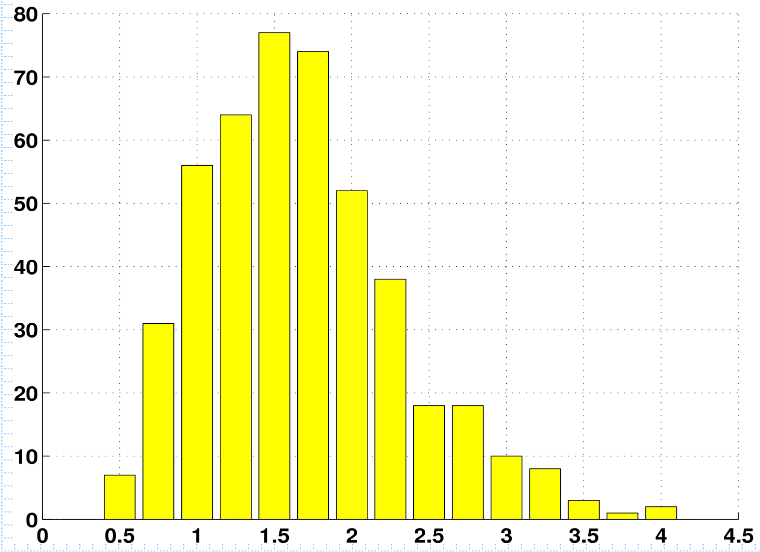


QPF, USA CONUS monthly

Time evolution of PCC for 6x6 tables

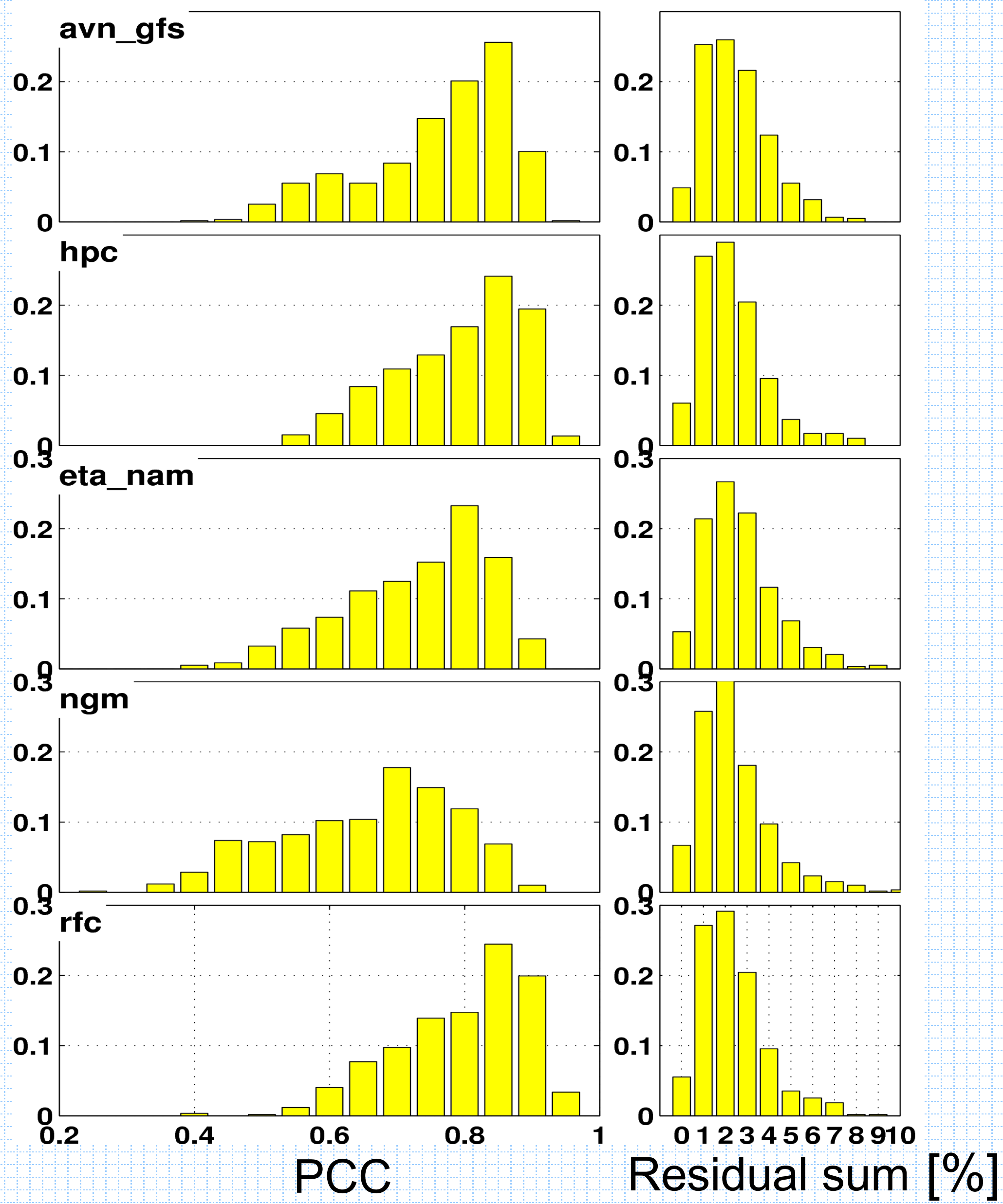


- Seasonal variation
- Slowly but constantly increasing trend
- Year to year variations



QPF, USA monthly tables, 2005-2009

PCC-s and residuals



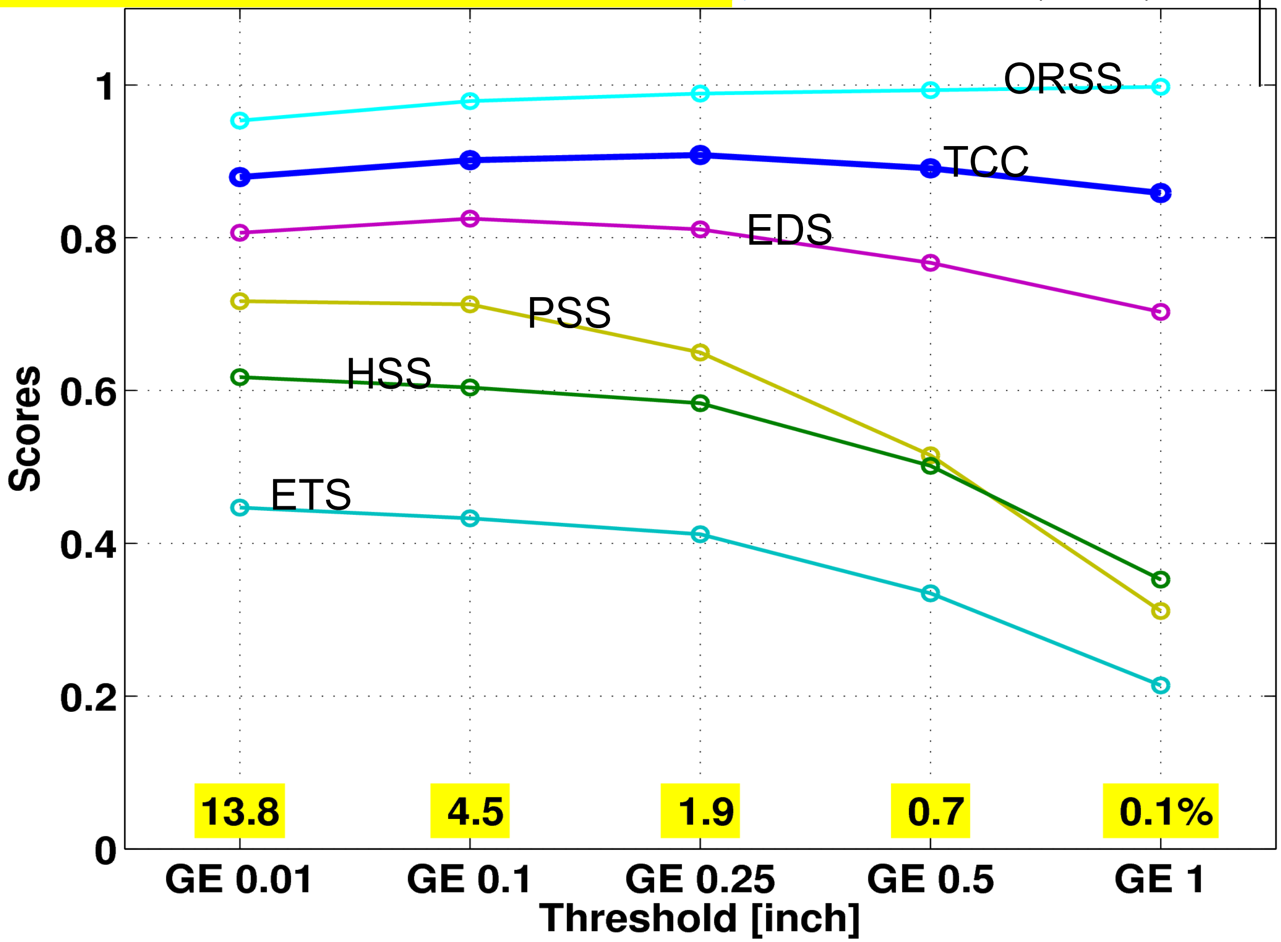
All 12 RFCs, together

QPF, USA CONUS 2008

Various scores for 2x2 tables from

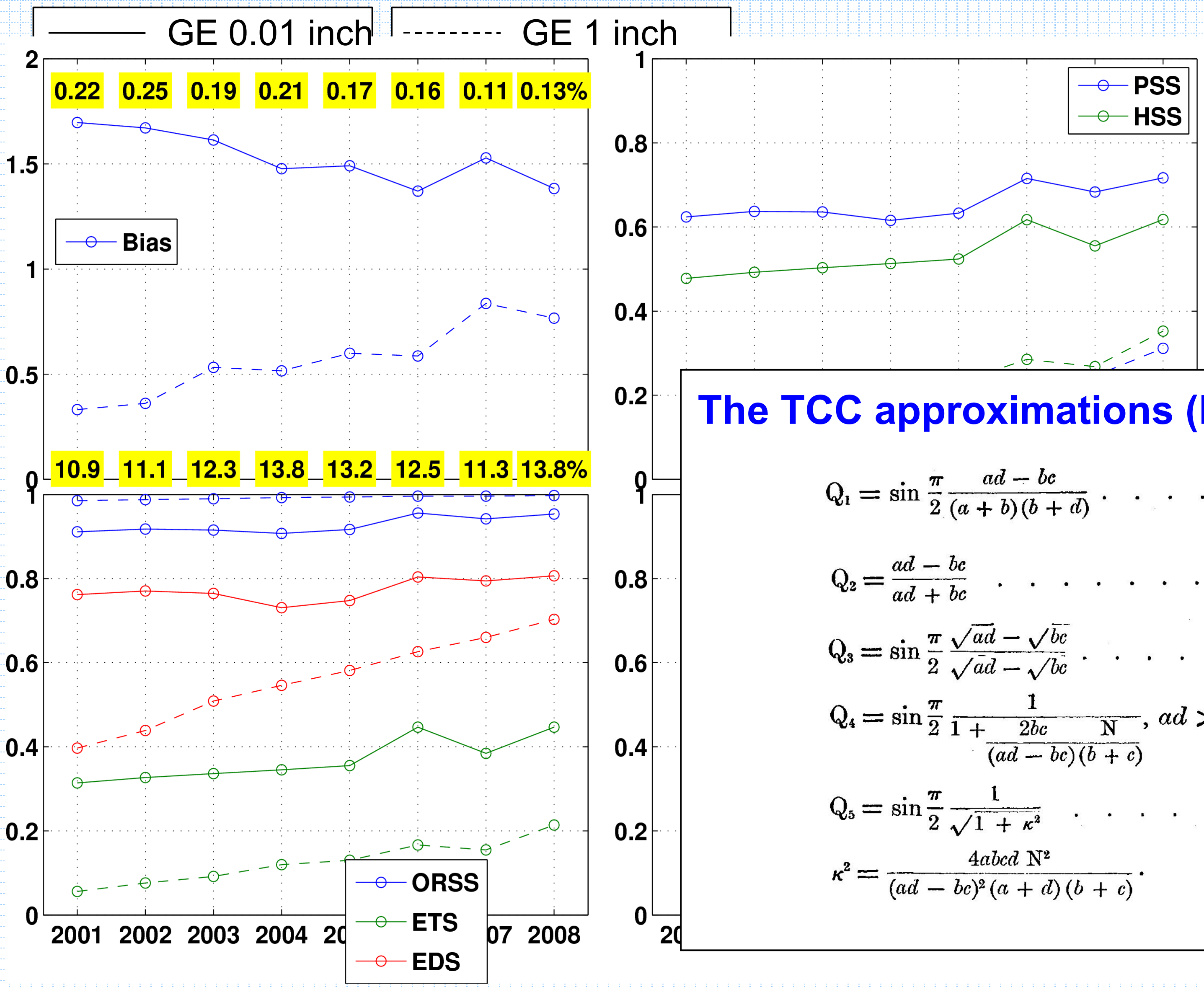
CAT	1	2	3	4	5	6	TOTALS
1	1041440	94627	9197	1396	168	9	1146837
2	32399	63464	22462	4335	664	34	123358
O 3	2216	10733	14304	6457	1284	60	35054
				6286	2526	166	16044
				2638	3020	490	7485
				289	720	521	1670
				21401	8382	1280	1330448

Dependence on the base rate



USA CONUS 2001-2008

Trends of various scores



The TCC approximations (Pearson,1900)

$$Q_1 = \sin\frac{\pi}{2}\,\frac{ad - bc}{(a+b)(b+d)} \quad . \ . \ . \ . \ . \ . \ . \ . \ . \quad \text{(lii.)}.$$

$$Q_2 = \frac{ad - bc}{ad + bc} \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \quad \text{(liii.)}.$$

$$Q_3 = \sin\frac{\pi}{2}\,\frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} - \sqrt{bc}} \quad . \ . \ . \ . \ . \ . \ . \ . \ . \quad \text{(lv.)},$$

$$Q_4 = \sin\frac{\pi}{2}\,\frac{1}{1 + \frac{2bc\ \ \ \ \ N}{(ad - bc)(b + c)}},\ ad > bc \quad . \ . \ . \quad \text{(lvi.)},$$

$$Q_5 = \sin\frac{\pi}{2}\,\frac{1}{\sqrt{1 + \kappa^2}} \quad . \ . \ . \ . \ . \ . \ . \ . \ . \quad \text{(lvii.)},$$

$$\kappa^2 = \frac{4abcd\ N^2}{(ad - bc)^2 (a + d)(b + c)}.$$

Many details not mentioned here, and especially so for the TCC, could be find in:

J. Juras and Z. Pasarić (2006):

Application of tetrachoric and polychoric correlation coefficients to forecast verification. *Geofizika*, **23**, 59-82.

(http://geofizika-journal.gfz.hr)

Thanks for your attention!!