Proper and equitable scores: a resolved dilemma

Zied Ben Bouallègue

7th International Verification Methods Workshop

Berlin, May 8-11, 2017

ECMWF

Definitions

In a nutshell:

- a **proper score** is based on a scoring rule consistent with the forecast interpretation
- an **equitable score** is a score for which all unskilful forecasts have the same expected score

Quote (1/2)

Proper Scores for Probability Forecasts Can Never Be Equitable
Jolliffe and Stephenson, 2008

“It would be ideal to have a verification score that is both proper and equitable, but …”

Elementary scoring matrix

Scoring matrix for forecast z, verification y, event threshold θ, penalty asymmetry α:

		$y \geq \theta$	
		1	0
$z \geq \theta$	1	0	α
	0	$1-\alpha$	0

Asymmetry

- penalty α in case of a false alarm
- penalty 1- α in case of a missed event

Elementary score for probabilistic forecasts

- defined for a given threshold θ and a given asymmetry α:

$$s_{\alpha,\theta}(z,y) = \begin{cases} 1-\alpha & \text{if} \quad y \geq \theta > z \\ \alpha & \text{if} \quad z \geq \theta > y \\ 0 & \text{otherwise} \end{cases}$$

- consistent scoring rule for **quantile forecasts** at probability level $1-\alpha$
- block unit for the definition of **proper probabilistic scores**

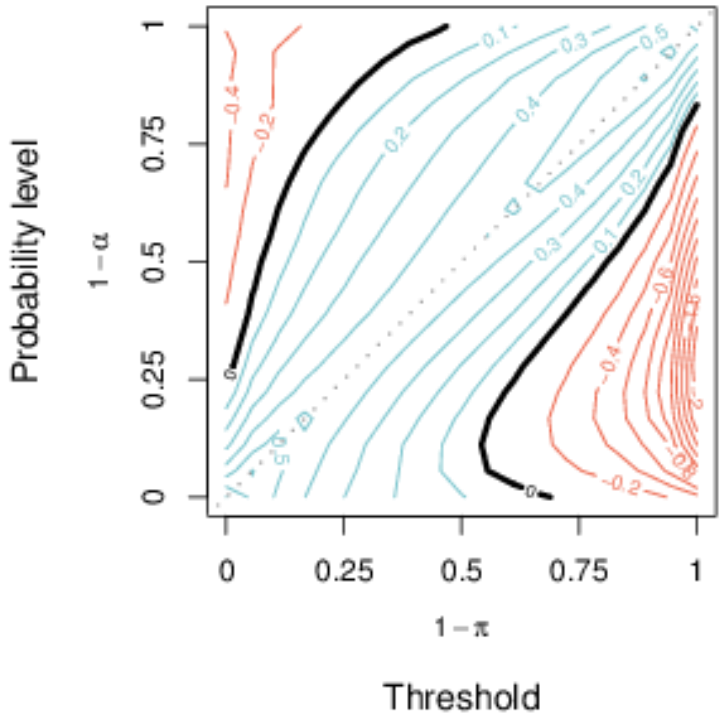
(Ehm et al. 2016)

- elementary skill score: forecast **economic value** for a user with a cost/loss ratio α

Forecast Skill Card

Elementary skill score as a function of :

- probability level $1 - \alpha$
- event base rate π

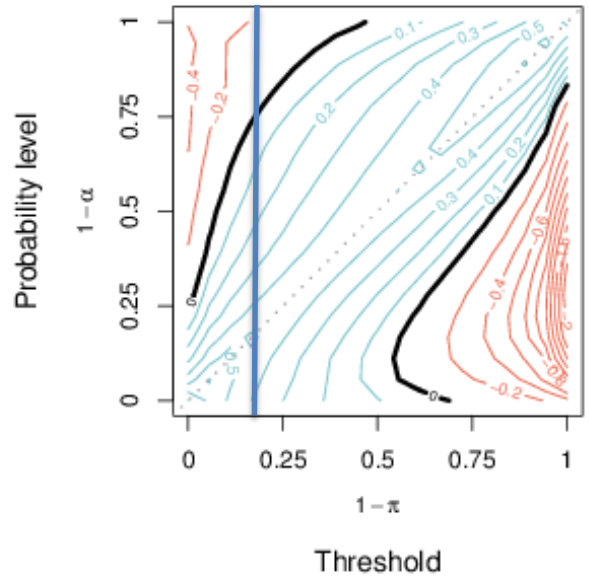


skill < 0
skill = 0
skill > 0

Forecast Skill Card of 2m-temprature
Europe, July 2016, 12UTC, day 5

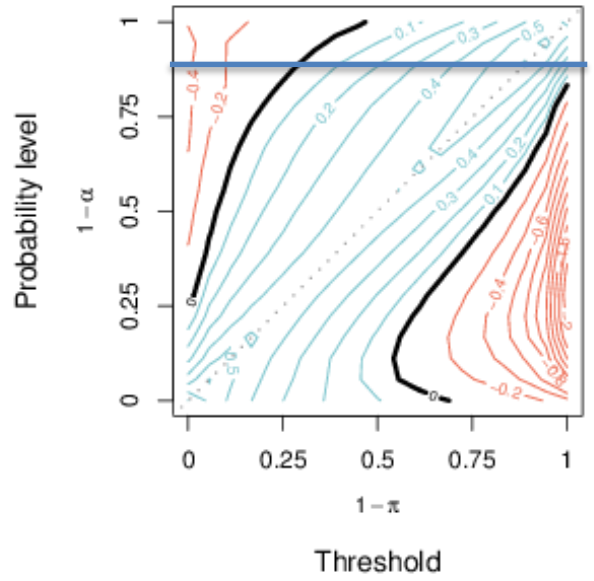
Summary scores

- Fixed event:
  integrate elementary (skill) score over a vertical line : **Brier (skill) score**



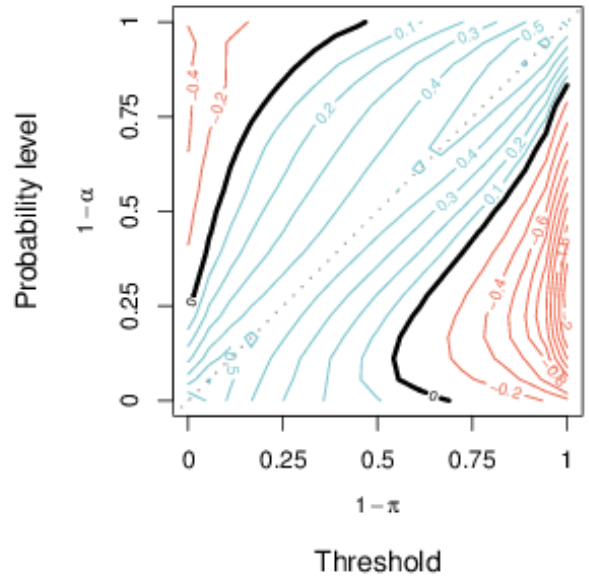
Summary scores

- Fixed event:
  integrate elementary (skill) score over a vertical line : **Brier (skill) score**
- Fixed probability level:
  integrate elementary (skill) score over a horizontal line : **Quantile (skill) score**



Summary scores

- Fixed event:
  integrate elementary (skill) score over a vertical line : **Brier (skill) score**
- Fixed probability level:
  integrate elementary (skill) score over a horizontal line : **Quantile (skill) score**
- All events and probability levels:
  integrate elementary (skill) score over the whole 2d card: **Continuous ranked prob. (skill) score**



Summary scores

- Fixed event:
  integrate elementary (skill) score over a vertical line : **Brier (skill) score**
- Fixed probability level:
  integrate elementary (skill) score over a horizontal line : **Quantile (skill) score**
- All events and probability levels:
  integrate elementary (skill) score over the whole 2d card: **Continuous ranked prob. (skill) score**

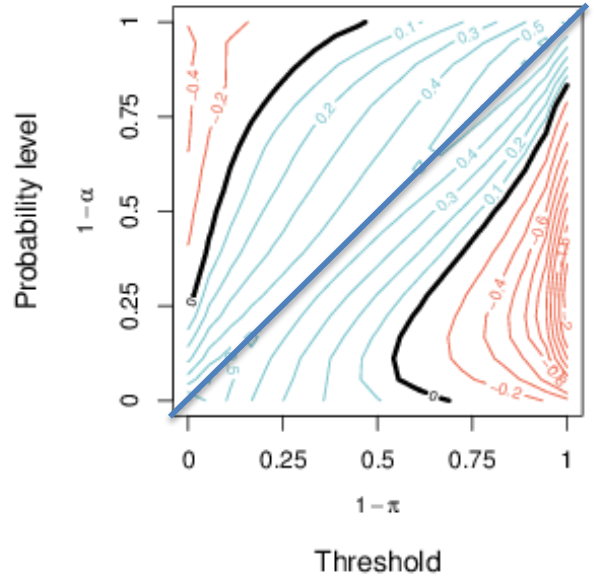
New approach:

- Fixed relationship between probability level and event base rate:

$$\alpha = \pi$$

integrate elementary (skill) score over the ascendant diagonal:

Diagonal (skill) score



Diagonal score(s)

Family of scores:

- Elementary diagonal scores :

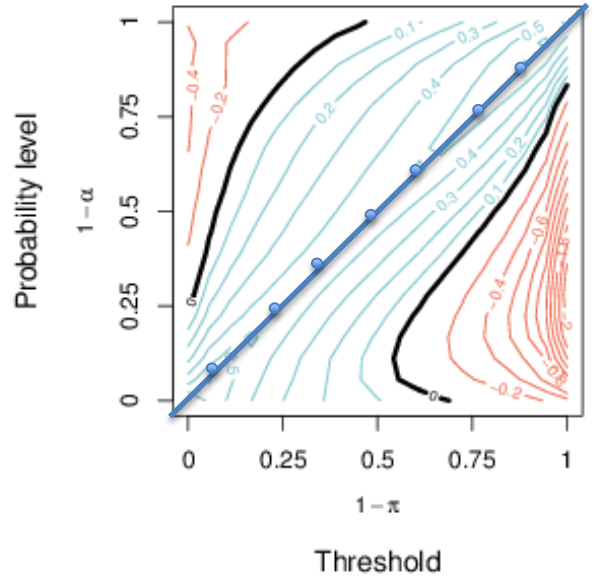
  expected score $\propto 1 - (H - F)$ (H= hit rate, F=false alarm rate)

- Ranked version : sum of elementary diagonal scores

- Continuous version : integral over all α

Proprieties:

- **Proper**: based on consistent scoring rules
- **Equitable**: same expected score when H=F



Synthetic dataset

Toy-model :

- Observation $Y \mid \mu \sim \mathcal{N}(\mu, \sigma^2) \quad with \quad \mu \sim \mathcal{N}(0, 1-\sigma^2)$
- Forecast $F \mid \mu \sim \mathcal{N}(\mu + b, \sigma^2) \quad with \quad \mu \sim \mathcal{N}(0, 1-\sigma^2)$

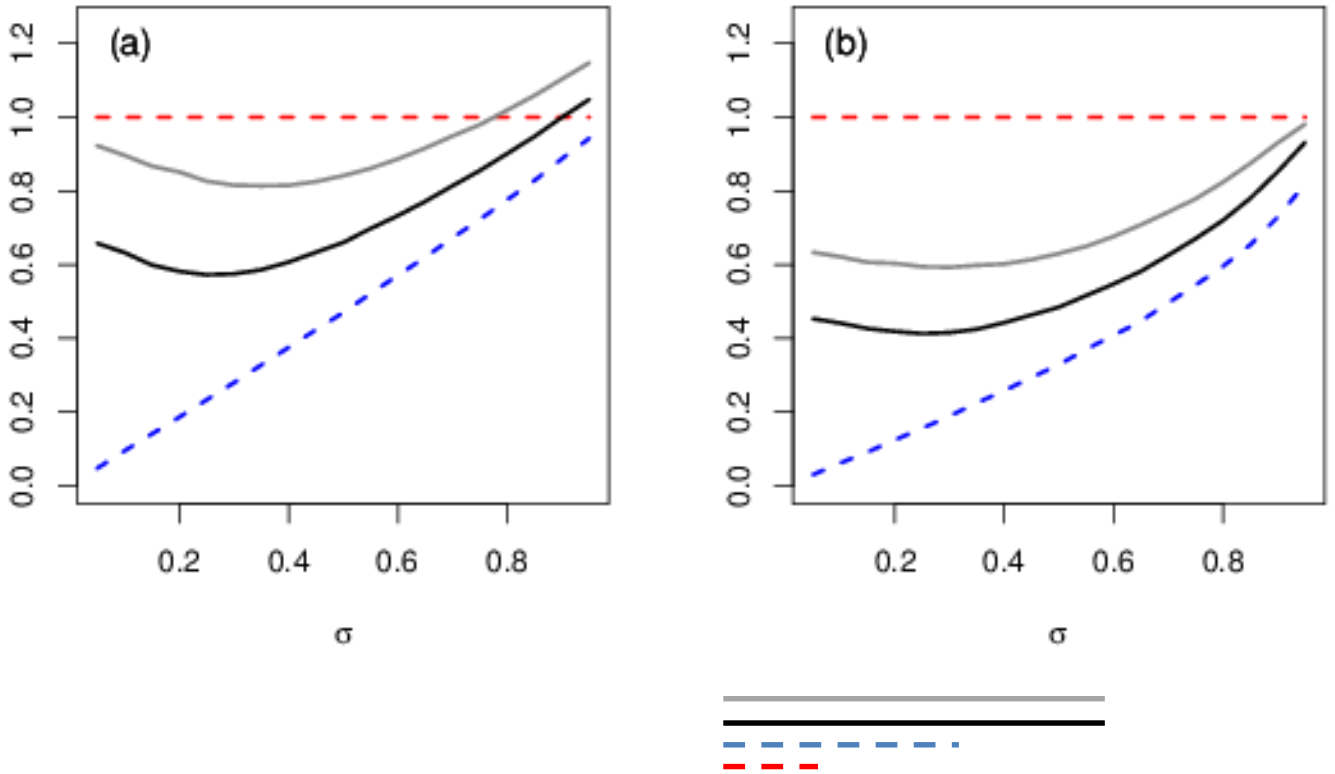
(Lerch et al. 2017)

b= forecast bias
σ governs predictability (ex: small σ, high signal-to-noise ratio)

- σ=0 perfect det. forecast
- σ=1 unconditional forecast

Score properties

Scores as a function of the signal-to-noise ratio:



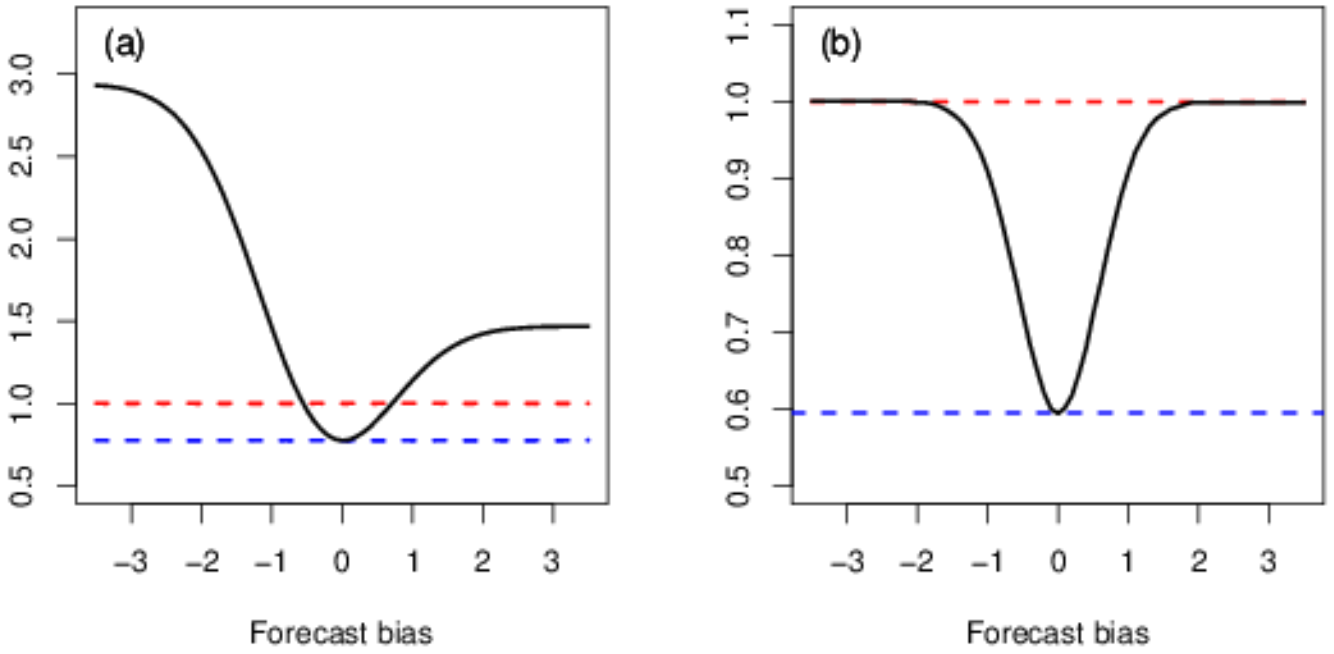
(a) norm. Brier score
(b) norm. diagonal score
for π =2/3

- unconditional forecast
- perfect forecast
- biased forecast (b=0.5)
- biased forecast (b=1)

- Decomposition of the scores into UNC - RES + REL
- Discrimination (resolution): necessary and sufficient condition for skill in **(b)**

Score properties

Scores as a function of the forecast bias:



(a) norm. Brier score
(b) norm. diagonal score
for π =2/3 σ=0.75

unconditional forecast
perfect forecast
biased forecast

- Score minimised when the forecast bias is zero (reliable forecast)
- Symmetry around unbiased forecast score and convergence towards uncond. forecast score in **(b)**

Quote (2/2)

The impenetrable hedge: a note on propriety, equitability and consistency
Jolliffe, 2008

“It is plausible that asymmetric scores may be required when cost and loss functions are taken into account, but to have highly **asymmetric scores determined solely by the base rate** seems less plausible”

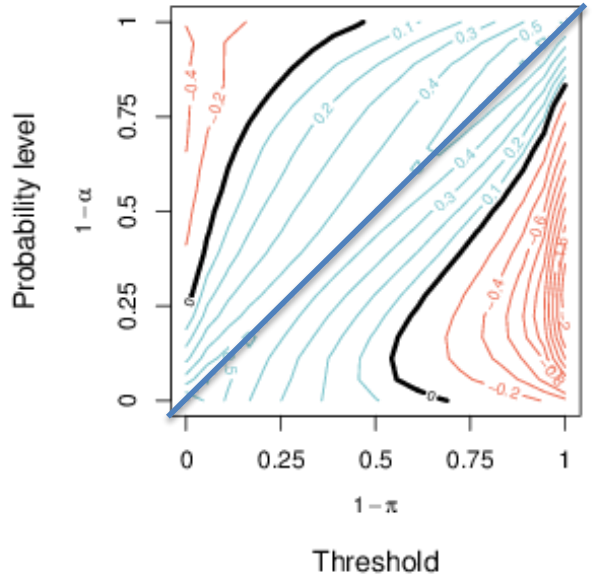
Diagonal score interpretation

Diagonal scores:

- Assume a fixed relationship between event base rate and cost-loss ratio ($\alpha = \pi$)
    - user's risk aversion increases with the intensity of the event under focus
    - appropriate for user with interest for high impact weather:
      rare events (small π) **and** potential high losses (small α)

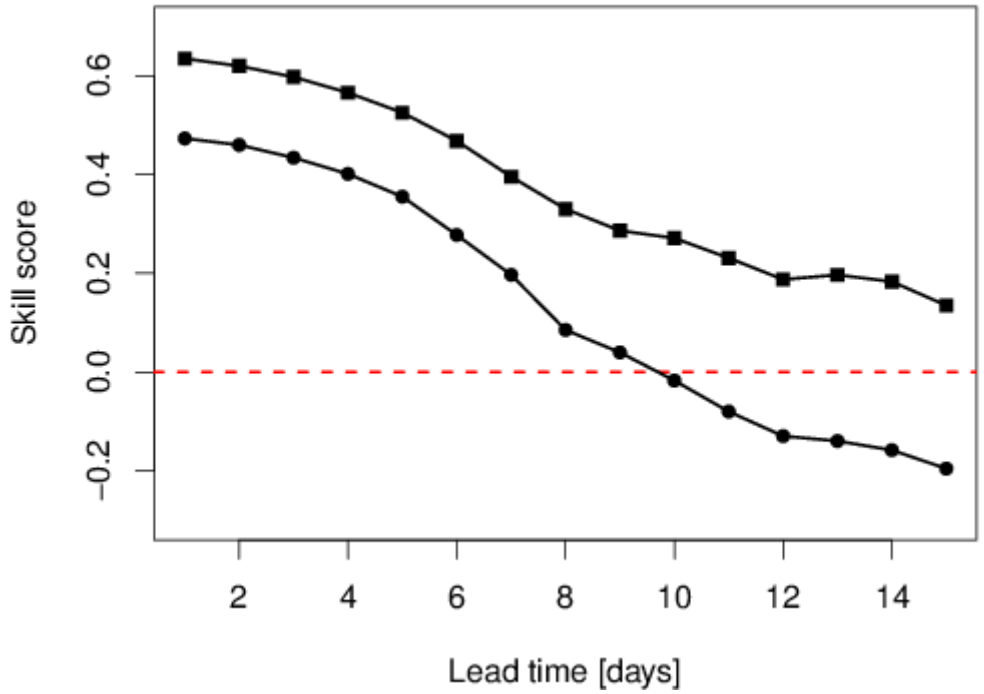
- Focus on the maximum discriminative ability of the system

(Manzato 2006)



2m-temperature

IFS ensemble performance as a function of the forecast lead time:



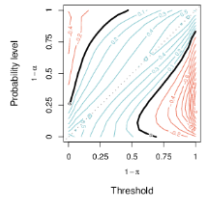
Skill scores based on
- Brier score
- Diagonal score

for π=2/3

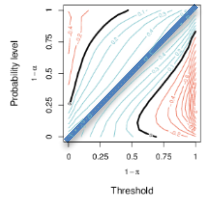
Summary

		$y \geq \theta$	
		1	0
$z \geq \theta$	1	0	α
	0	$1-\alpha$	0

- **Elementary score** : block score unit for the definition of consistent scoring rules



- **Forecast skill card** : elementary skill scores as the event and probability level varies



- **Diagonal score** : new summary score with *interesting* interpretations and properties

References

Ehm W., T. Gneiting, A. Jordan and F. Krüger, 2016. Of Quantiles and Expectiles: Consistent Scoring Functions, Choquet Representations, and Forecast Rankings, J.R. Statist. Soc. B, 136,1505-1510.

Jolliffe I.T. and D.B. Stephenson, 2008. Proper scores for probability forecasts can never be equitable. Mon. Weather Rev., 136,1505-1510.

Jolliffe I.T., 2008. The impenetrable hedge: a note on propriety, equitability and consistency. Met. Apps,15, 25-29.

Lerch S., T. L. Thorarinsdottir, F. Ravazzolo and T. Gneiting, 2017. Forecaster's Dilemma: Extreme Events and Forecast Evaluation, Statist. Sci., 32, 106-127.

Manzato A., 2006. A Note On the Maximum Peirce Skill Score, Wea. Forecasting, 22, 1148-1154.