Receiver Operating Characteristic (ROC) Curves

Tilmann Gneiting
Heidelberg Institute for Theoretical Studies (HITS) and
Karlsruhe Institute of Technology (KIT)

Peter Vogel
CSL Behring, Marburg

Eva-Maria Walz
KIT and HITS

TIGGE–S2S Workshop, ECMWF, Reading

3 April 2019

Karlsruhe Institute of Technology

Heidelberg Institute for Theoretical Studies

HITS

WAVES TO WEATHER

Outline

1. Probability forecasts
2. Receiver operating characteristic (ROC) curves
3. ROC — The movie
4. Recommendations for forecast verification

Gneiting, T. and P. Vogel (2018). **Receiver operating characteristic (ROC) curves.** Preprint, `arXiv:1809.04808`.

Vogel, P., P. Knippertz, A. H. Fink, A. Schlueter, and T. Gneiting (2018). **Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa.** *Weather and Forecasting*, 33, 369–388.

Walz, E.-M. (2018). **A generalization of ROC curves.** Master thesis, Faculty of Mathematics, Karlsruhe Institute of Technology.

Outline

1. Probability forecasts
2. Receiver operating characteristic (ROC) curves
3. ROC — The movie
4. Recommendations for forecast verification

Vogel, P., P. Knippertz, A. H. Fink, A. Schlueter, and T. Gneiting (2018). **Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa.** *Weather and Forecasting*, 33, 369–388.

Probabilistic forecasts

Probabilistic forecasts take the form of predictive probability distributions over future quantities or events

Have become state of the art in many scientific disciplines and application domains, including but not limited to

- Meteorology
- Hydrology
- Renewable energy
- Medicine
- Economics
- Finance

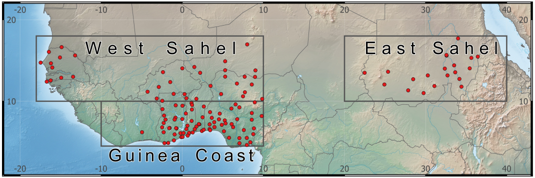
Simplest case is a probability forecast for a binary event, typically defined in terms of a threshold, such as

- Precipitation occurrence
- Flooding
- Extreme wind speed
- Cancer diagnosis
- Recession
- Credit default

Probability forecasts

Probability forecasts specify a predictive probability for a binary event of interest, typically defined in terms of a threshold

Example (Vogel et al. 2018): 24-hour Probability of Precipitation (PoP) forecasts from the ECMWF ensemble system over northern tropical Africa



We consider the binary event of precipitation occurrence at a threshold of 0.2 mm, for both observations and forecasts

Outline

1. Probability forecasts
2. Receiver operating characteristic (ROC) curves
3. ROC — The movie
4. Recommendations for forecast verification

Gneiting, T. and P. Vogel (2018). **Receiver operating characteristic (ROC) curves.** Preprint, `arXiv:1809.04808`.

Receiver operating characteristic (ROC) curve

Receiver (or Relative) Operating Characteristic (ROC) curves are ubiquitously used to evaluate probability forecasts:

- According to the Web of Science, myriads (!) of scientific papers employ ROC curves
- A supplementary headline score at ECMWF is based on AUC for the extreme forecast index (EFI) of (binarized) 10m wind speed
- The WMO mandates the use of ROC curves for verifying (binarized) long range temperature forecasts (SVSLRF)

Essentially, the ROC curve plots the hit rate (HR) versus the false alarm rate (FAR) as the predictor threshold x varies, where

$$\mathrm{HR}(x) = \frac{\mathrm{TP}(x)}{\mathrm{TP}(x) + \mathrm{FN}(x)} \quad \text{and} \quad \mathrm{FAR}(x) = \frac{\mathrm{FP}(x)}{\mathrm{FP}(x) + \mathrm{TN}(x)}$$

The Area Under the ROC Curve (AUC) is a positively oriented measure of predictive ability

Area Under the ROC Curve (AUC)

The Area Under the ROC Curve (AUC) is a positively oriented measure of predictive ability

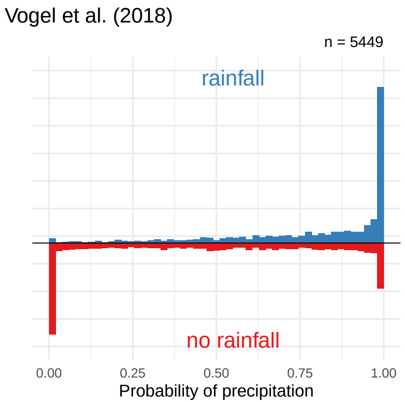
- Appealing interpretation as the probability that a (randomly chosen) predictor value under an event is larger than a (randomly chosen) predictor under a non-event:

$$\mathrm{AUC} = \mathbb{P}(X' > X \mid Y' = 1, Y = 0)$$

- $\mathrm{AUC} = (D + 1)/2$ in terms of Somers' D (Somers 1962)
- $\mathrm{AUC} = 1/2$ and $D = 0$ for a useless predictor that is independent of the binary event of interest
- $\mathrm{AUC} = 1$ and $D = 1$ for a perfect predictor

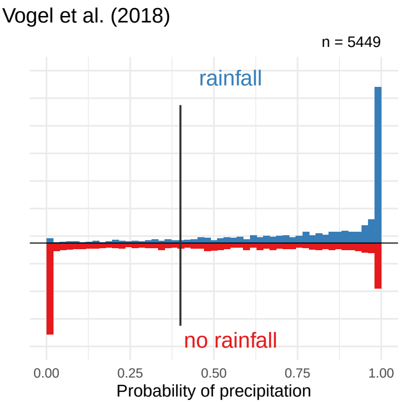
Example

ROC curve and AUC for 24-hour Probability of Precipitation (PoP) forecasts (at a threshold of 0.2 mm) from the ECMWF ensemble over West Sahel (Vogel et al. 2018)



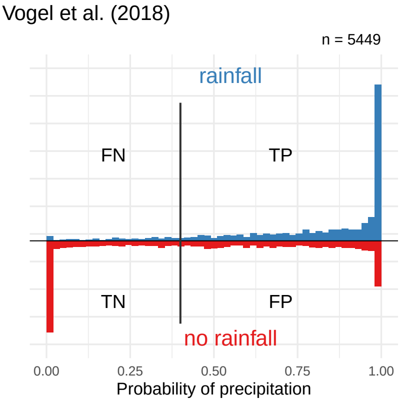
Example

ROC curve and AUC for 24-hour Probability of Precipitation (PoP) forecasts (at a threshold of 0.2 mm) from the ECMWF ensemble over West Sahel (Vogel et al. 2018)



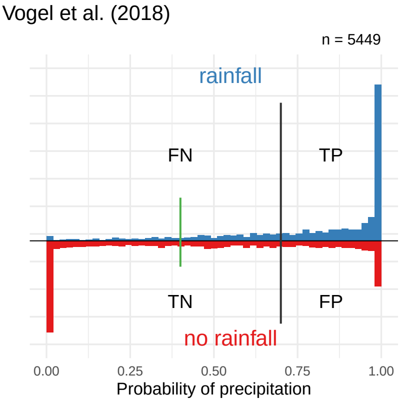
Example

ROC curve and AUC for 24-hour Probability of Precipitation (PoP) forecasts (at a threshold of 0.2 mm) from the ECMWF ensemble over West Sahel (Vogel et al. 2018)



Example

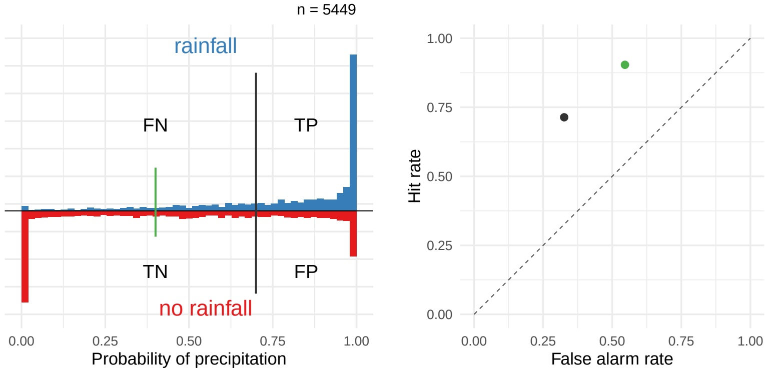
ROC curve and AUC for 24-hour Probability of Precipitation (PoP) forecasts (at a threshold of 0.2 mm) from the ECMWF ensemble over West Sahel (Vogel et al. 2018)



Example

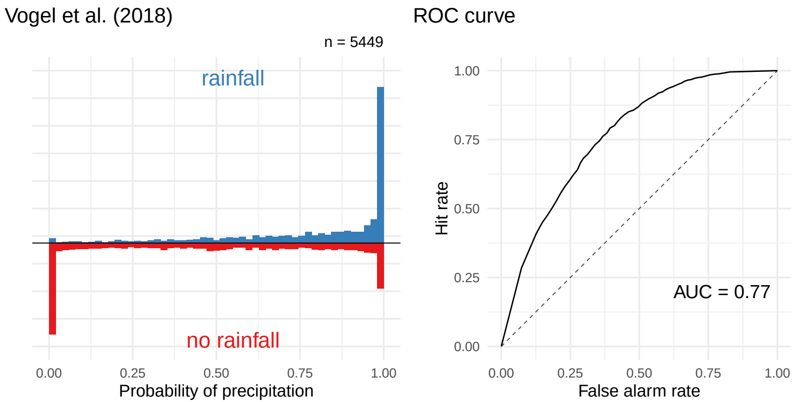
ROC curve and AUC for 24-hour Probability of Precipitation (PoP) forecasts (at a threshold of 0.2 mm) from the ECMWF ensemble over West Sahel (Vogel et al. 2018)

Vogel et al. (2018)



Example

ROC curve and AUC for 24-hour Probability of Precipitation (PoP) forecasts (at a threshold of 0.2 mm) from the ECMWF ensemble over West Sahel (Vogel et al. 2018)



A formal approach to ROC curves

Formal setting:

X real-valued predictor
Y binary outcome
$\mathbb{P}$ joint distribution of (X, Y)

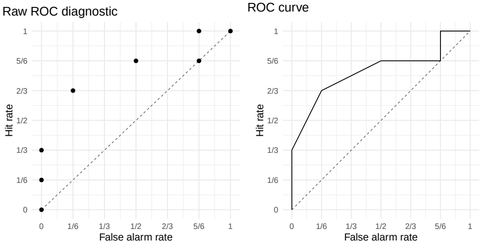
$$F_0(x) = \mathbb{P}(X \leq x \mid Y = 0)$$
$$F_1(x) = \mathbb{P}(X \leq x \mid Y = 1)$$

The raw ROC diagnostic is the set of all points of the form

$$(\mathrm{FAR}(x), \mathrm{HR}(x)) \in [0,1] \times [0,1]$$

threshold $x \in \mathbb{R}$, $\mathrm{FAR}(x) = 1 - F_0(x)$, $\mathrm{HR}(x) = 1 - F_1(x)$

The ROC curve is the linearly interpolated raw ROC diagnostic



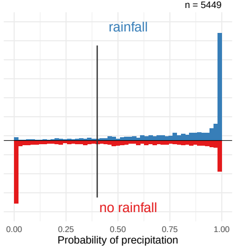
Properties of ROC curves and AUC

Interpretation as function: For continuous, strictly increasing F_0 and F_1,

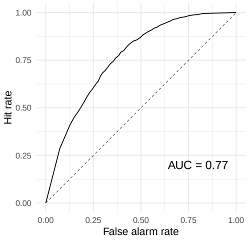
$$R(\alpha) = 1 - F_1(F_0^{-1}(1-\alpha)), \quad \alpha = \mathrm{FAR}(x) \in [0,1]$$

Ensuing math fact: Characterization of ROC curves

Vogel et al. (2018)



ROC curve



Properties of ROC curves and AUC

Interpretation as function: For continuous, strictly increasing F_0 and F_1,

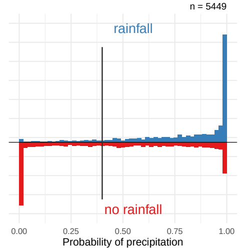
$$R(\alpha) = 1 - F_1(F_0^{-1}(1 - \alpha)), \quad \alpha = \mathrm{FAR}(x) \in [0, 1]$$

Ensuing math fact: Characterization of ROC curves

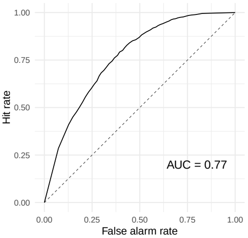
Invariance of ROC curves and AUC under

- changes in class proportions
- strictly increasing transformations of the predictor X

Vogel et al. (2018)



ROC curve



Properties of ROC curves and AUC

Interpretation as function: For continuous, strictly increasing F_0 and F_1,

$$R(\alpha) = 1 - F_1(F_0^{-1}(1-\alpha)), \quad \alpha = \mathrm{FAR}(x) \in [0,1]$$

Ensuing math fact: Characterization of ROC curves

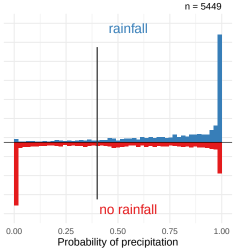
Invariance of ROC curves and AUC under

- changes in class proportions
- strictly increasing transformations of the predictor X

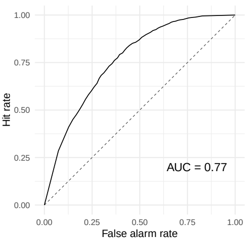
Consequence (Mason and Graham 2002; Kharin and Zwiers 2003): ROC curves and AUC

- do not consider calibration
- nor economic value,
- and apply to real-valued predictors X on arbitrary scales

Vogel et al. (2018)



ROC curve

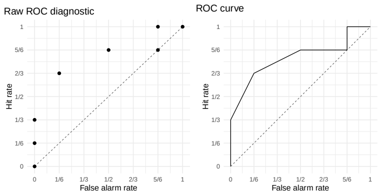


Crucial role of concavity

Pesce et al. (2010): The use of non-concave ROC curves is "irrational" and "unethical when applied to medical decisions"

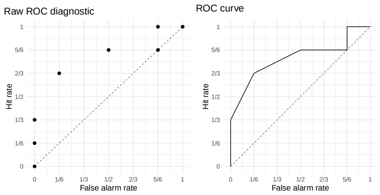
Crucial role of concavity

Pesce et al. (2010): The use of non-concave ROC curves is “irrational” and “unethical when applied to medical decisions”



Crucial role of concavity

Pesce et al. (2010): The use of non-concave ROC curves is "irrational" and "unethical when applied to medical decisions"



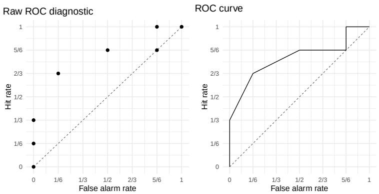
Theorem: The following statements are equivalent:

(a) The conditional event probability $\mathrm{CEP}(x) = \mathbb{P}(Y = 1 \mid X = x)$ is nondecreasing in the decision threshold x

(b) The ROC curve is concave

Crucial role of concavity

Pesce et al. (2010): The use of non-concave ROC curves is "irrational" and "unethical when applied to medical decisions"



Theorem: The following statements are equivalent:

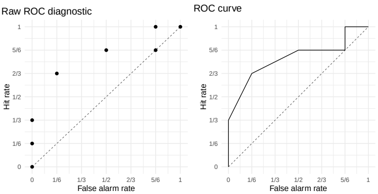
(a) The conditional event probability $\mathrm{CEP}(x) = \mathbb{P}(Y = 1 \mid X = x)$ is nondecreasing in the decision threshold x

(b) The ROC curve is concave

Bottom line: If we believe that the conditional event probability increases with the predictor value, we should insist on using concave ROC curves only!

Crucial role of concavity

Pesce et al. (2010): The use of non-concave ROC curves is “irrational” and “unethical when applied to medical decisions”



Theorem: The following statements are equivalent:

(a) The conditional event probability $\mathrm{CEP}(x) = \mathbb{P}(Y = 1 \,|\, X = x)$ is nondecreasing in the decision threshold x

(b) The ROC curve is concave

Bottom line: If we believe that the conditional event probability increases with the predictor value, we should insist on using concave ROC curves only!

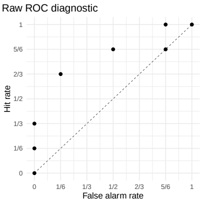
But: Non-concave ROC curves occur inevitably, reflecting noise in the data.

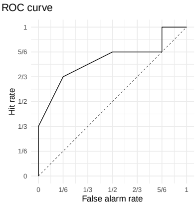
Enforcing non-decreasing CEPs and concave ROC curves

The classical pool-adjacent-violators (PAV) algorithm (Ayer et al. 1955)

- turns X into a modified predictor X^{PAV} with non-decreasing conditional event probabilities (CEPs),
- morphs a ROC curve into its concave hull, and
- improves the AUC value

X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	
Y	0	1	0, 0	0, 0, 1	0, 1, 1	1	1	
FAR(x)	5/6	5/6	1/2	1/6	0	0	0	
HR(x)	1	5/6	5/6	2/3	1/3	1/6	0	
CEP(x)	0	1	0	1/3	2/3	1	1	AUC $= 112/144$



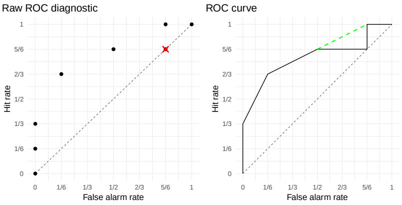


Enforcing non-decreasing CEPs and concave ROC curves

The classical pool-adjacent-violators (PAV) algorithm (Ayer et al. 1955)

- turns X into a modified predictor X_{PAV} with non-decreasing conditional event probabilities (CEPs),
- morphs a ROC curve into its concave hull, and
- improves the AUC value

X_{PAV}	x_1	$x_{2,3}$	x_4	x_5	x_6	x_7	
Y	0	1, 0, 0	0, 0, 1	0, 1, 1	1	1	
FAR(x)	5/6	1/2	1/6	0	0	0	
HR(x)	1	5/6	2/3	1/3	1/6	0	
CEP(x)	0	1/3	1/3	2/3	1	1	AUC $=$ 116/144



Outline

1. Probability forecasts
2. Receiver operating characteristic (ROC) curves
3. ROC — The movie
4. Recommendations for forecast verification

Walz, E.-M. (2018). **A Generalization of ROC Curves**. Master thesis, Faculty of Mathematics, Karlsruhe Institute of Technology.

Motivation

Despite their ubiquitous use and popularity, ROC curves and AUC are subject to a major limitation:

The target variable Y needs to be binary

For decades, researchers have sought a generalization that allows for real-valued target variables

Hernández-Orallo (2013, p. 3395): It is "questionable whether a similar graphical representation [. . .] can be figured out"

In the Master thesis project of Eva-Maria Walz (2018), we have made major steps towards the desired generalization

ROC movie and universal ROC (uROC) curve

Thresholding the target variable yields a sequence of (classical) ROC curves, which can be visualized in a ROC movie

Assigning weights to these curves and averaging accordingly results in a universal ROC (uROC) curve

- Invariant under strictly monotone transformations

The Area Under the ROC Movie (AUM) is a positively oriented measure of predictive ability

- Appealing interpretation as (weighted) probability that predictor and outcome are concordant
- For continuous variables, $\mathrm{AUM} = (\rho_S + 1)/2$ in terms of Spearman's ρ_S
- $\mathrm{AUM} = 1/2$ and $\rho_S = 0$ for a useless predictor; $\mathrm{AUM} = 1$ and $\rho_S = 1$ for a perfect predictor
- In the case of a binary outcome, ROC movie, uROC curve and AUM reduce to the classical ROC curve and AUC, respectively

Outline

1. Probability forecasts
2. Receiver operating characteristic (ROC) curves
3. ROC — The movie
4. Recommendations for forecast verification

Kharin, V. and F. Zwiers (2003). **On the ROC score of probability forecasts.** *Journal of Climate*, 16, 4145–4150.

Vogel, P., P. Knippertz, A. H. Fink, A. Schlueter, and T. Gneiting (2018). **Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa.** *Weather and Forecasting*, 33, 369–388.

Three crucial insights . . . illustrated on African precipitation

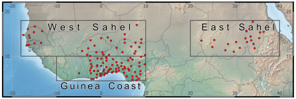
Crucial insight 1 ROC curves should be concave . . . if we believe that larger forecasts are indicative of larger outcomes!

Crucial insight 2 ROC curves and AUC assess potential predictive ability (only) . . . so for evaluating probability forecasts they should be accompanied by reliability diagrams and Murphy diagrams

Crucial insight 3 Appealing generalizations of ROC curves and AUC to real-valued target variables are feasible . . . premiere of ROC movie, uROC curve and AUM to follow!

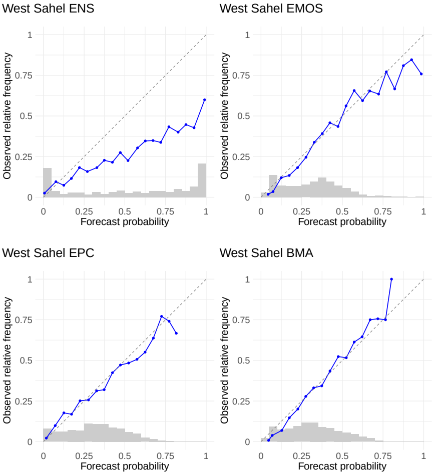
Example

24-hour precipitation forecasts over the West Sahel region in northern tropical Africa in monsoon season 2014 (Vogel et al. 2018)



Competing Probability of Precipitation (PoP) forecasts:

- **ENS** Raw ECMWF ensemble
- **EPC** Extended Probabilistic Climatology
- **EMOS** Calibrated by Ensemble Model Output Statistics
- **BMA** Calibrated by Bayesian Model Averaging

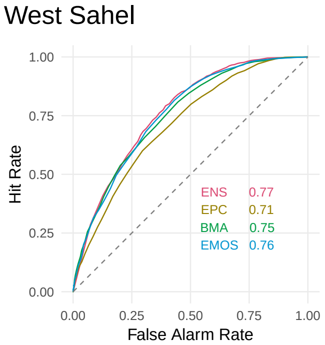


ROC curves should be concave

If we believe that larger forecasts are indicative of larger observations, we should only be using concave ROC curves

Isotony (of the predictor) and concavity (of the ROC curve) can be enforced with the pool-adjacent-violators (PAV) algorithm

Free lunch — the transition to the concave hull benefits AUC as well!

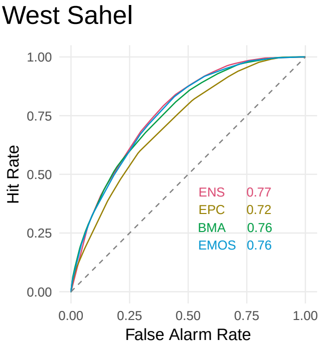


ROC curves should be concave

If we believe that larger forecasts are indicative of larger observations, we should only be using concave ROC curves

Isotony (of the predictor) and concavity (of the ROC curve) can be enforced with the pool-adjacent-violators (PAV) algorithm

Free lunch — the transition to the concave hull benefits AUC as well!



ROC curves and AUC assess potential predictive ability

Invariance under strictly monotone transformations has stark implications:

- ROC curves and AUC can be used to assess the predictive ability of just any real-valued predictor
- However, for probability forecasts, calibration and actual economic value get ignored
- To be used in concert with reliability diagrams and Murphy diagrams

Murphy diagram

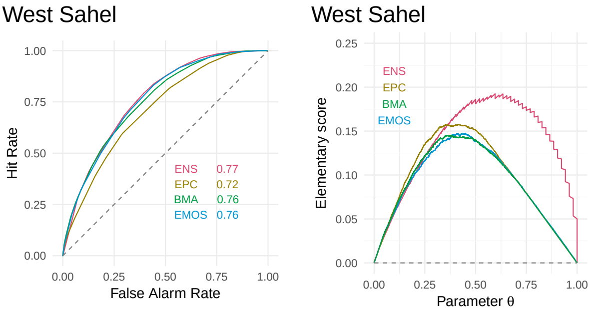
- Every proper scoring rule is a mixture of elementary scores

$$S_\theta(p, y) = \begin{cases} \theta, & y = 0, p > \theta, \\ 1 - \theta, & y = 1, p \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

- A Murphy diagram plots the mean elementary scores of competing forecasts as a function of $\theta \in (0, 1)$
- Covers all economic scenarios simultaneously and eliminates the need to choose a proper scoring rule (Murphy 1977; Ehm et al. 2016)

ROC curves and Murphy diagrams . . . illustrated

- ROC curves and AUC assess potential predictive ability, i.e., actual predictive ability subsequent to postprocessing
- Murphy diagrams visualize actually incurred (normalized) cost for a binary decision maker with expense ratio $\theta/(1-\theta)$



ROC movie, uROC curve and AUM: The premiere

24-hour quantitative precipitation forecasts (ECMWF ensemble mean) over the West Sahel region in northern tropical Africa in monsoon season 2014

Both the predictor and the target variable are real-valued now