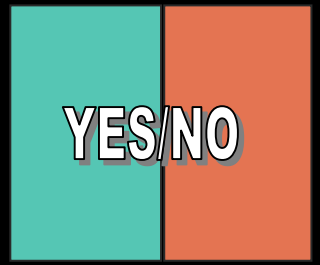
Categorical scores



Marion Mittermaier

MPEDTP Virtual Summer School

22 June 2021

Topics covered

- The Murphy and Winkler (1987) framework, expanding on concepts such as joint/marginal distributions.
- Contingency table
- selecting metrics that use all elements of the CT; focus on independent and complementary scores
- Different kinds of thresholds (strengths and weaknesses); false skill
- Strengths and weaknesses of categorical metrics
- Double penalty and km-scale modelling
- Multi-categorical contingency tables, Gerrity score.
- Newer scores such as SEEPS that incorporate climatology
- Skill scores: types of reference forecasts (strengths and weaknesses)
- Different observations of the same thing producing conflicting results

Basic concepts

- ***Categorical:*** only one of a set of possible events will occur
- Categorical forecast ***does not contain expression of uncertainty***
- There is typically a one-to-one correspondence between the forecast values and the observed values.
- The **simplest possible situation is a 2x2 case** or verification of a categorical yes/no forecast: 2 possible forecasts (yes/no) and 2 possible outcomes (event observed/event not observed)

Binary (dichotomous) events

Yes/No, presence/absence

Examples

- It’s foggy
- There’s a frost
- It’s raining
- There’s a thunderstorm (in a given area)

Underlying continuous variable

- Visibility
- Air / ground temperature
- Rainfall accumulation
- Lightning flash rate

Dichotomization

Advantages

- Easier to explain and present
- User-meaningful thresholds
  - e.g. aviation, wind power

Disadvantages

- Lose information
- Threshold(s) must be chosen
  - Can be hard, may be arbitrary

- **Simpler to analyze?**
- ***I'm afraid not!***

The Murphy & Winkler framework 1

Based on the **joint distribution** of forecasts and observations.

Designed to work for a **simple yes/no** (e.g. rain, no rain), **multiple categories** (e.g. clear, partly cloudy, overcast) and even something **probabilistic** (e.g. 60% chance of rain tomorrow).

Let $p(f,x)$ denote the **joint probability** of the forecasts f and the observations x, which contains information about the forecast, the observations and the relationship between the two.

(Note: a yes/no categorical forecast is a special case of a probability forecast where a probability of 1 is assigned to a particular event.)

The Murphy & Winkler framework 2

Because we want the best forecasts possible we would like a *high proportion of forecasts to be close to the observat*ions and a *low proportion to be far away*.

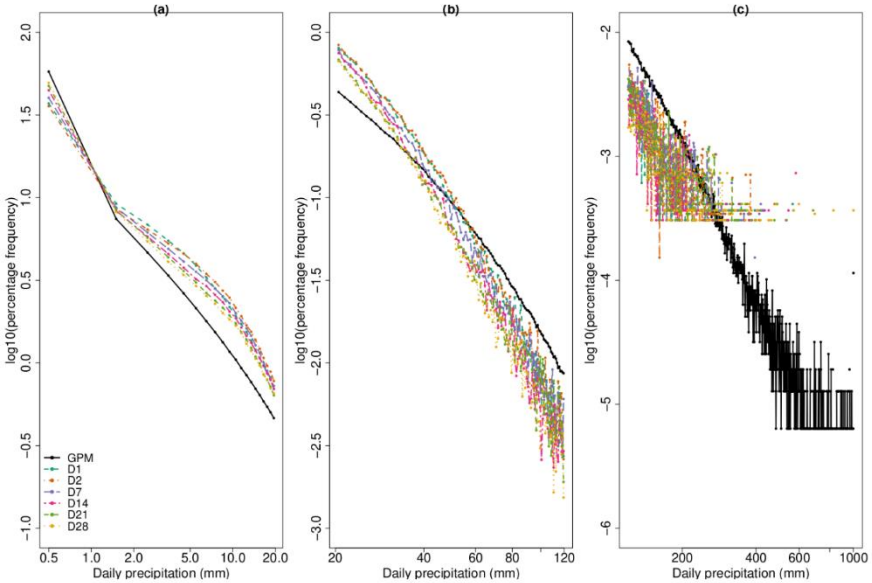
To learn more about the forecasts the **joint distribution can be factorised** into a combination of *marginal* and *conditional* distributions to give **two perspectives**.

1) **Calibration-refinement**: $p(f, x) = p(x \mid f) \, p(f)$ (how often different observations have been made for a given forecast)
2) **Likelihood-base rate**: $p(f, x) = p(f \mid x) \, p(x)$ (how often are different forecasts made before it is observed and $p(x)$ is the **base rate (frequency of observed occurrence)**

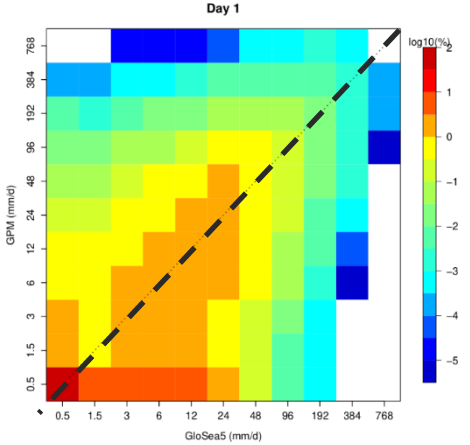
The two factorisations are related through Bayes' theorem:

$$p(x \mid f) = [\, p(x) \, p(f \mid x) \,] \, / \, p(f)$$

Example of marginal and joint distributions



Marginal distributions of daily rainfall for different lead times



Joint (binned, multi-category) distribution of daily rainfall between GPM and D1 forecasts (if there was perfect correspondence all the values would be in the bins along the diagonal

Output from MET tool GridDiag

The scatter plot and joint distributions

Plot of **matched** (observation, forecast) **pairs**

Perfect when data points fall on 1:1 line

The humble scatter plot is totally under-rated and often dismissed for being "too simple" ….

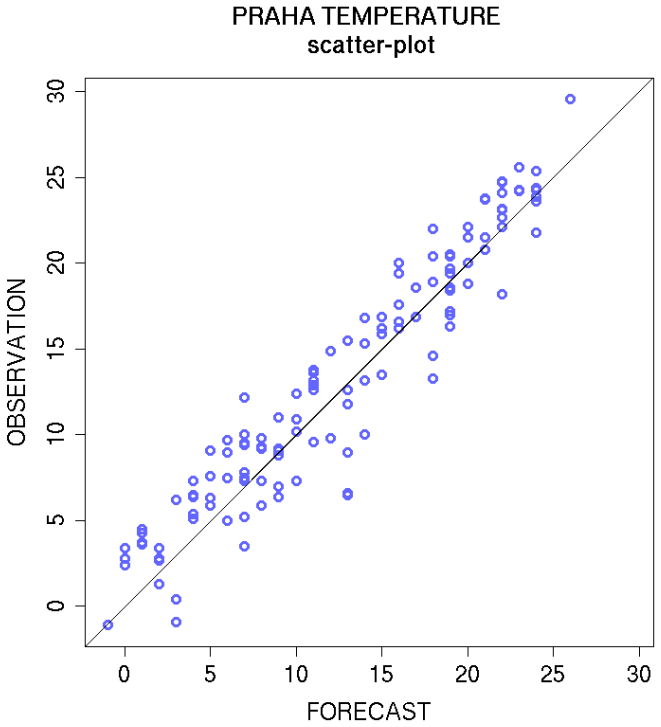
However, it is incredibly useful for examining the nature of your dataset:

Is there a bias?

Are there outliers?

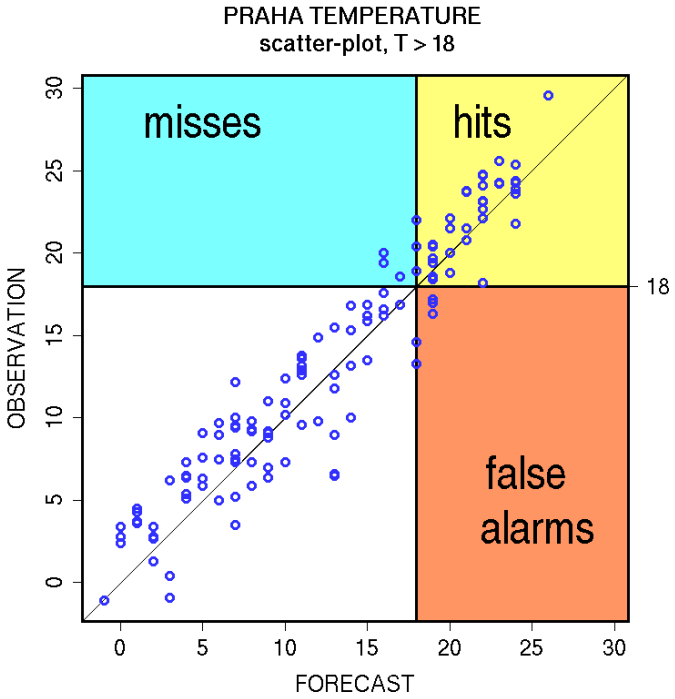
Is there a linear association? How strong? Does it change with magnitude?

What happens in the tails?

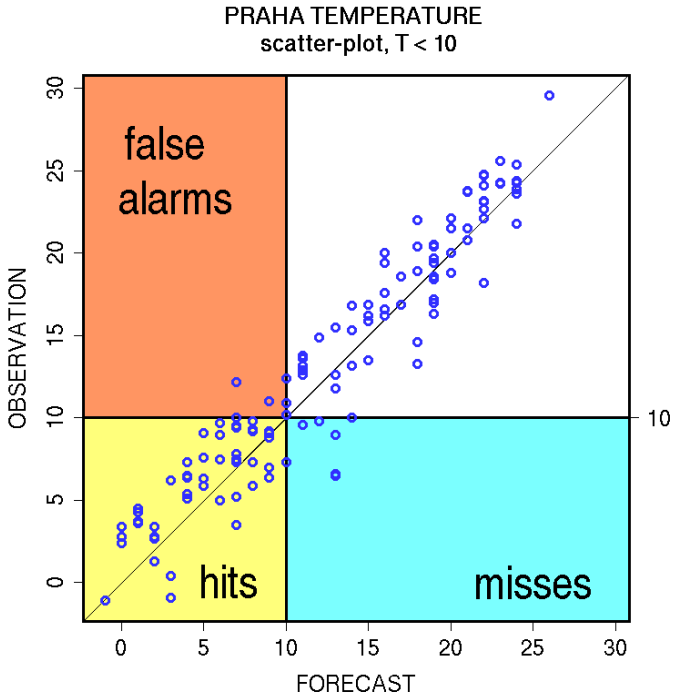


The link between a scatter plot and a contingency table

Does the forecast detect correctly temperatures above 18 degrees ?



Does the forecast detect correctly temperatures below 10 degrees ?



How do these concepts translate to contingency tables?

Marginal probability: sum of column or row divided by the total sample size
For example: the marginal probability of a yes forecast is:
p_x=Pr(X=1) = 100/2800 = 0.03

Tornado forecast	Tornado Observed		
	yes	no	Total fc
yes	30	70	100
no	20	2680	2700
Total obs	50	2750	2800

Sum of rows

Total sample size

How do these concepts translate to contingency tables?

Joint probability: represents the intersection of *two* events in a cross-tabulation table.
For example: the joint probability of a yes forecast and a yes observed:

$$P_{x,y}=Pr(X=1,Y=1) = 30/2800 = 0.01$$

Tornado forecast	Tornado Observed		
	yes	no	Total fc
yes	30	70	100
no	20	2680	2700
Total obs	50	2750	2800

Met Office

Standard notation

Computed as follows from a dataset F and O and an event threshold th:

Hits = *a* = (F >= th) & (O >= th)

False alarms = *b* = (F>= th) & (O < th)

Misses = *c* = (F < th) & (O >= th)

Correct rejs = *d* = (F < th) & (O < th)

***Q**: Why might the categorical evaluation of warnings be problematic?*

Event forecast?	Event observed?		
	Yes	No	Total:
Yes	*hits*	*false alarms*	*events forecast*
No	*misses*	*correct rejections*	*non-events forecast*
Total:	*events observed*	*non-events observed*	*all events (sample size)*

Event forecast?	Event observed?		
	Yes	No	Total
Yes	*a*	*b*	*a + b*
No	*c*	*d*	*c + d*
Total	*a + c*	*b + d*	*n = a + b + c + d*

Need three measures to completely describe the joint probability distribution (n is the denominator)

Need 3 quantities to fully describe joint distribution

- There are a plethora of contingency-table based scores to compute. The easiest non-textbook reference is https://community.wmo.int/jwgfvr-verification.

- *You can choose to compute them all.* It's easy enough to do. However, **most of them are NOT independent**. To **fully describe the joint distribution** there are **only two options** for obtaining independent information. Any categorical analysis should always include these:

1. Base rate *s*, Probability of Detection (POD) and Probability of False Detection (POFD)

***or* :**

2. Base rate *s*, frequency bias *B* and an overall accuracy measure

- The base rate *s* is a property of climatology so cannot be altered by the forecast system → base rate *s* (and sample size *n*) characterise the observations
- Another *two* measures are needed to characterise the forecast system

Independent measures

- Base rate

  *s = events **observed** / all events*
  = (a + c) / n

- Frequency bias

  *B = events **forecast** / events **observed***
  = (a + b) / (a + c)

- Probability of detection (hit rate)

  POD = hits / events observed
  = a / (a + c)

  Range [0, 1]

- Probability of false detection (false alarm rate)

  POFD = false alarms / non-events observed
  = b / (b + d)

  Range [0, 1]

Not to be confused with

- False alarm **ratio**

  Range [0, 1]

  *FAR = false alarms / events **forecast***
  = b / (a + b)

Not independent!

Met Office

Peirce skill score
(C.S. Peirce, 1884)

Using a performance metric that is built on H *and* F can be useful. Here, we give *TWO* examples….

- Known by many names: (true skill statistic, Hanssen & Kuiper's discriminant, Kuiper's performance index, Youden's index, risk difference …)

$$PSS = H - F = \frac{(ad - bc)}{(b + d)(a + c)}$$ Range [−1, 1]

- Has a simple definition (based on *H* and *F*) and good properties.
- Penalizes misses and false alarms equally
- Can be extended to multiple categories.

Met Office

Rare events

Note: It took various authors 3 attempts at getting this right! Earlier versions (EDS, SEDS) have been largely superseded/improved upon – no reason to use them.

(S)EDI have better properties and though EDI has a simpler definition and is easier to understand, there is no reason not to use SEDI in practice (v. similar value for rare events, sensible also when event definition switched i.e. for v common events).

- As **the base rate *s* tends to zero**, most performance measures (PC, PSS, TS, ETS, HSS) tend to a meaningless fixed value (1 for PC, 0 for others).
  – They are ***degenerate* for rare events**
- A measure that avoids this is the (symmetric) extremal dependence index

$$(S)EDI = \frac{\ln(F) - \ln(H)\ [+ \ln(1-H) - \ln(1-F)]}{\ln(F) + \ln(H)\ [+ \ln(1-H) + \ln(1-F)]}$$

- (S)EDI is independent of the base rate *s*

Not all rare events are extreme though most extreme events are relatively rare

Interpretation: is the value of my score good or bad?

- Given a score of *x*, is that good or not? For the basics (base rate and bias) it is relatively easy. But is a POD of 0.2 good? What about a SEDI of 0.7?
- The **“goodness”** (for a user) can be difficult to establish, dependent on the user and potentially **highly subjective.**
- If calculated values are near the bounds it can be considered “easier” but any values near the bounds should also *make* you ask yourself some (other) scientific questions (i.e. relating the metric back to the meteorology/science). Here a **broader understanding of your forecasts and your observations is essential**. This can often be improved by considering the datasets spatially and/or temporally, i.e. not just in some aggregated sense.
- **Looking at a range of scores can help**, provided you understand each of the metrics, what they measure and their strengths and weaknesses. Otherwise it can confuse….
- **Most contingency based scores are not measures of skill, just a measure of performance or accuracy. Skill scores** are generally easier to interpret because they are *relative to the accuracy of another forecast*.

Verification history: Finley's tornadoes (1884)

Tornado forecast	Tornado Observed		
	yes	no	Total fc
yes	30	70	100
no	20	2680	2700
Total obs	50	2750	2800

PC = (a+d)/n

PC=(30+2680)/2800= 96.8%

POD = 30/50 = 60%
POFD = 70/100 = 70%
B = 100/50 = 2

Then Gilbert (1884) showed soon after that …

Tornado forecast	Tornado Observed		
	yes	no	Total fc
yes	0	0	0
no	50	2750	2800
Total obs	50	2750	2800

PC=(2750+0)/2800= 98.2%

POD = 0 = 0%
POFD = 0 = 0%
B = 0/50 = 0

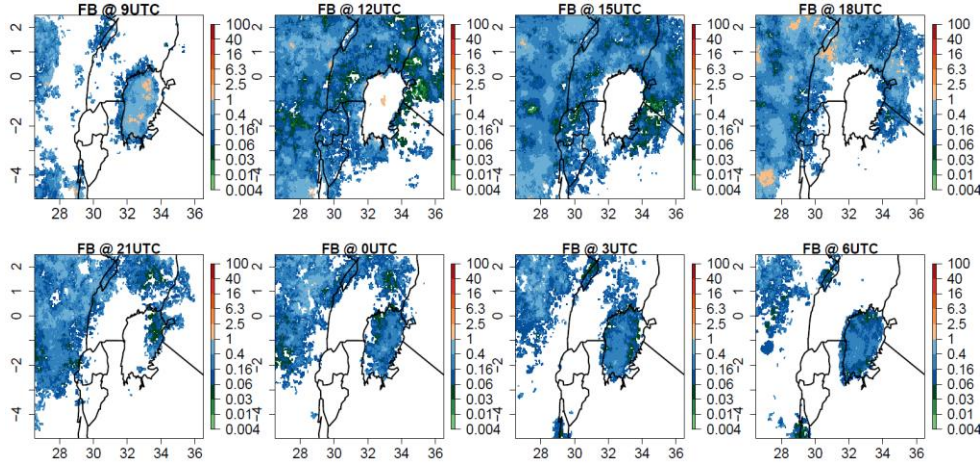
So, if you forecast nothing you do better?! Erm, no.

Met Office

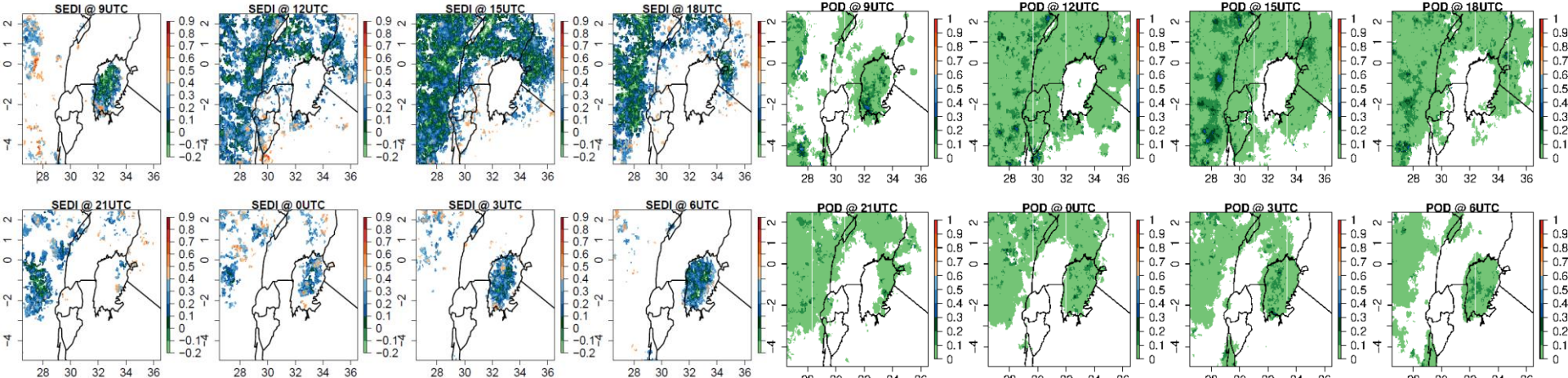
Scores for the occurrence of any lighting from a 4.4 km model and gridded lightning data over Lake Victoria region for period ~192 day period broken down in space and time.

Practical question:

1. What sampling effects may come into play when computing scores on a grid? (i.e. a time aggregate at a point)
2. Comparing the SEDI and the POD, would you say the forecast is skilful?
3. If you aggregated the domain and produced a line graph, could this change your interpretation?



From Mittermaier et al. (2021a)

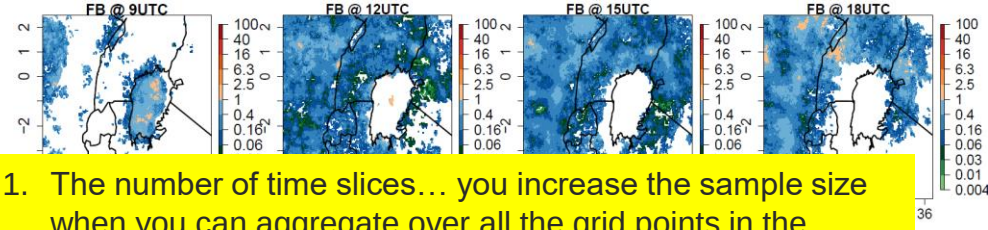


Met Office

Scores for the occurrence of any lighting from a 4.4 km model and gridded lightning data over Lake Victoria region for period ~192 day period broken down in space and time.

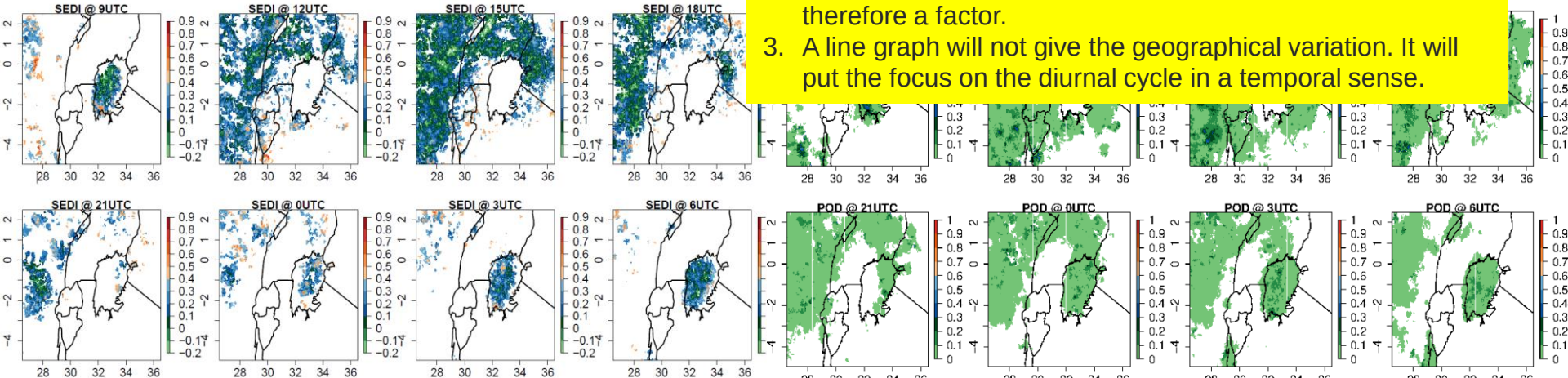
Practical question:

1. What sampling effects may come into play when computing scores on a grid? (i.e. a time aggregate at a point)
2. Comparing the SEDI and the POD, would you say the forecast is skilful?
3. If you aggregated the domain and produced a line graph, could this change your interpretation?



1. The number of time slices… you increase the sample size when you can aggregate over all the grid points in the domain.
2. POD values appear quite uniform in space, and nothing larger than 0.3. SEDI values are larger but much more variable. There are locations where there is a POD but no SEDI… this is because of the use of logarithms in SEDI. Any 0s will mean the score is undefined. Sample size is therefore a factor.
3. A line graph will not give the geographical variation. It will put the focus on the diurnal cycle in a temporal sense.

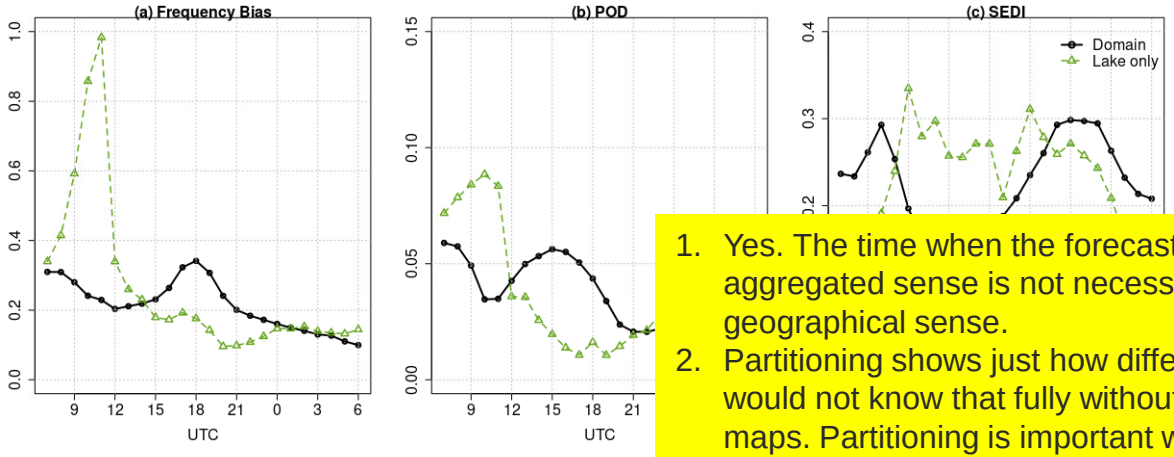
From Mittermaier et al. (2021a)



Scores for the occurrence of any lighting from a 4.4 km model and gridded lightning data over Lake Victoria region for period ~192 day period broken down in space and time.

Practical question part 2:

Here is an aggregate of the scores over the domain and over 192 days as a function of time of day.



3. Do these line graphs change your interpretation from the spati
4. What does the partitioning do? Why could this be important m
5. How should scores be aggregated? Why?

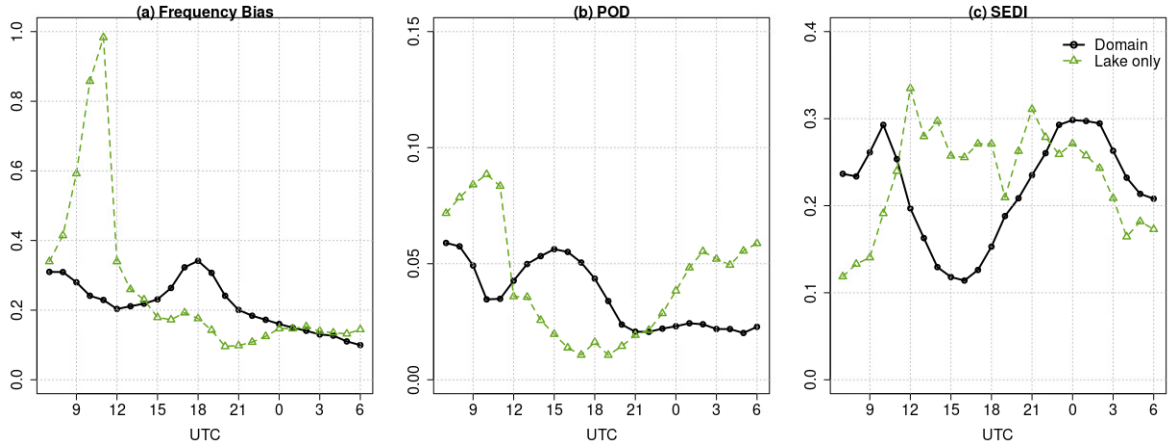
1. Yes. The time when the forecast is more skill and a spatially aggregated sense is not necessarily when it was in a geographical sense.
2. Partitioning shows just how different the lake is, but you would not know that fully without looking at the spatial maps. Partitioning is important when you think there are very different signals in space or time which may be masked, providing a false impression of performance.
3. Categorical scores should not be aggregated.. The contingency tables are the atomic component for aggregation and/or stratification.

Met Office

Scores for the occurrence of any lighting from a 4.4 km model and gridded lightning data over Lake Victoria region for period ~192 day period broken down in space and time.

Practical question part 2:

Here is an aggregate of the scores over the domain and over 192 days as a function of time of day.



3. Do these line graphs change your interpretation from the spatial maps?
4. What does the partitioning do? Why could this be important more generally?
5. How should scores be aggregated? Why?

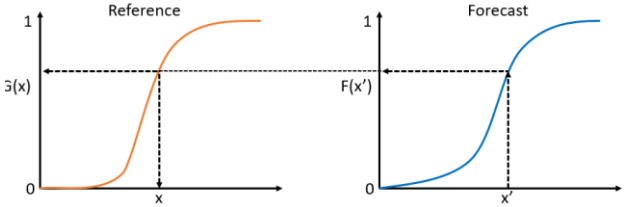
Measuring skill by constructing a skill score

$$SS = \frac{Score - Score_{ref}}{Score_{perf} - Score_{ref}}$$

- **Measure the forecast accuracy with respect to the accuracy of a reference forecast: positive values = skill; negative values = no skill**
- Difference between the score and a reference forecast score, normalized by the score obtained for a perfect forecast minus the reference forecast score
- *In theory there is no limit to which scores are used to construct a skill score* so long as you know what the perfect score is and you can compute a score for your chosen reference forecast.

Choosing thresholds wisely

- Thresholds can be highly arbitrary and/or driven by user requirements. If the latter, think about the implications very carefully.
- Some rules to consider:
  - Choose thresholds that **suit the data** (what is the underlying distribution?)
  - **Is there a bias?** Choose metrics that are insensitive to the bias or choose to transform your data to remove/mitigate against it, e.g. via some form of quantile mapping. In this case the bias has to be monitored separately.



  - **Use long-term climatological thresholds or climatologically relevant thresholds**. This is particularly important when attempting to aggregate over regions so that the thresholds are relevant to each location.

Types of reference forecasts

To measure skill you need to compare to something that can be considered a forecast for the observed event.

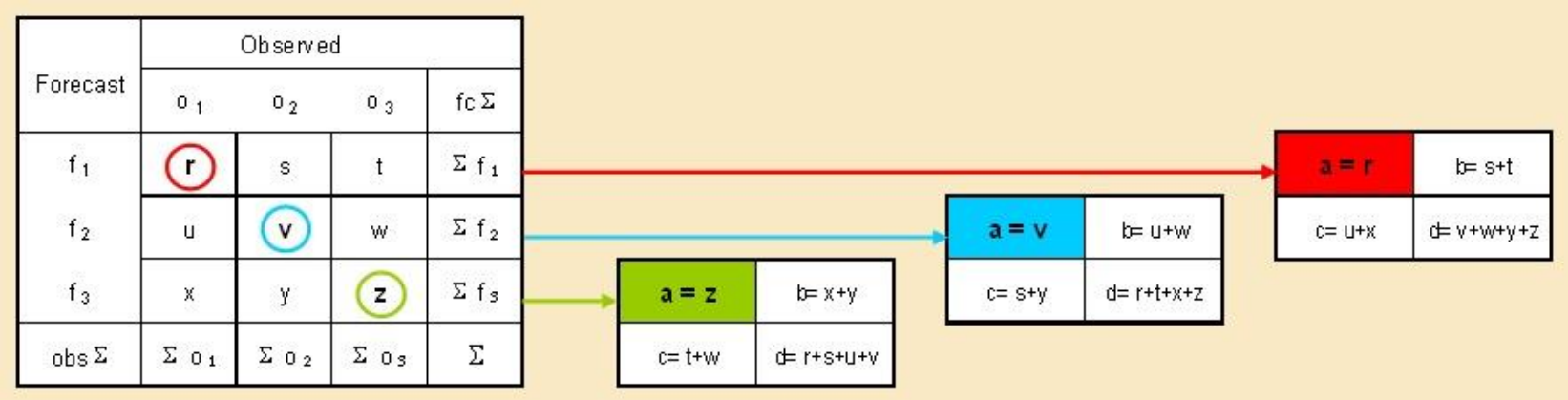
Type	Example	Properties
Random	Equitable Threat Score	• Well understood statistical benchmark • Not physically meaningful
Persistence	Constructed skill score	• Measure of predictability (predictability is low when persistence is a poor forecast) • Show value added by running NWP model
Sample climate a posteriori	Constructed skill score	• One step further removed than persistence, i.e. smoothed • Retains predictability element due to regime dependence
Long-term climatology a priori	Constructed skill score, extremes	• Easiest reference to beat, smoothest • Care required i.t.o. representativeness, pooling issues, climate change trends

Met Office

Multi-category events

- Equitability (or at least asymptotic equitability is important.
- These can become extremely technical and there can be many pitfalls, i.e. proceed with caution.
- Most commonly created from splitting continuous variable at multiple thresholds
  - 6-hour precip: 0.5, 1, 4 millimetres
  - Total cloud amount: clear, partly cloudy, cloudy (defined by some bounds)
  - Visibility 200, 1000, 4000 metres
- Most often analysed as a set of 2×2 tables by dichotomizing at each threshold in turn
  - Giving three 2×2 tables in each of above case

Multi-category events



Generalized version of HSS and PSS - measure of improvement over random forecast

$$HSS = \frac{\left\{ \sum p(f_i, o_i) - \sum p(f_i)p(o_i) \right\}}{\left\{ 1 - \sum p(f_i)p(o_i) \right\}}$$

$$KSS = \frac{\left\{ \sum p(f_i, o_i) - \sum p(f_i)p(o_i) \right\}}{\left\{ 1 - \sum (p(f_i))^2 \right\}}$$

Hedging: best thought of as a sensitive to the bias, e.g. the score can be improved through over forecasting, i.e. not forecasting the true frequency

Multi-category events: single overall scores

- For **ordered** categories (as formed by splitting continuous variables), most common is the **Gerrity score**
- **LEPSCAT** is an alternative
  - But less sensitive to skill trends and more vulnerable to *hedging*
- **SEEPS (Stable Equitable Error in Probability Space)** often preferable for *three* ordered categories (but can be extended to more)
  - Sensitive to trends in forecast skill
  - Designed for precipitation but there is no reason why it can't be used for other quantities provided a climatology can be computed.

Gerrity score

- **Recommended score for 3 x 3 tables** with many desirable properties.
- Uses a scoring matrix which is a tabulation of the reward or penalty for every forecast/observation outcome represented by the contingency table.
- Uses sample probabilities, not those on which the categories are based (i.e., not 1/3, 1/3, 1/3 for a 3 x 3 table)
- Provides consistency in terms of any underlying correlations and from categorisation to categorisation.
- The score is equitable, does not depend on the forecast distribution, does not reward conservatism, utilizes off diagonal information in the contingency table, and penalizes larger errors more.

$$GSS = \sum_{i=1}^{3} \sum_{j=1}^{3} p_{ij} s_{ij}$$

The scoring matrix is given by

$$s_{ii} = \frac{1}{2}\left(\sum_{r=1}^{i-1} a_r^{-1} + \sum_{r=i}^{2} a_r \right)$$

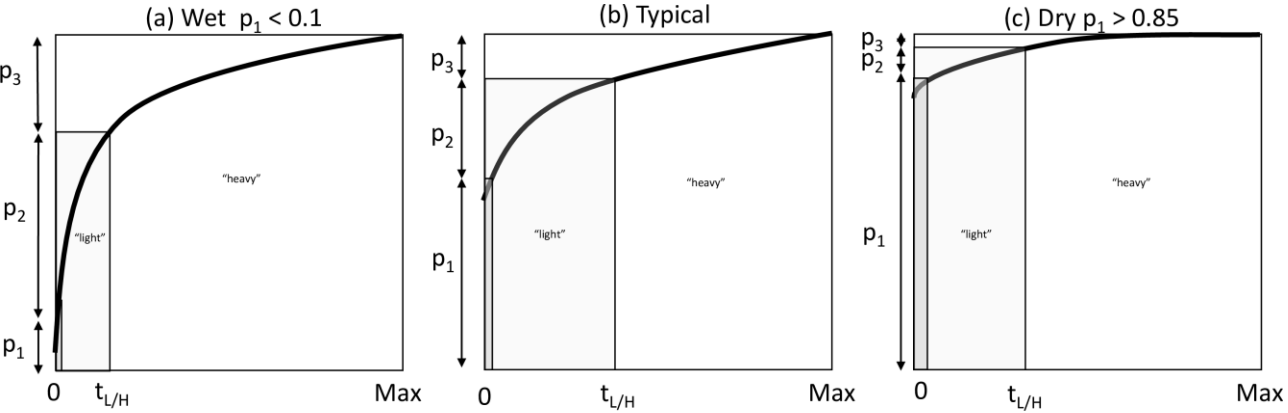
$$s_{ij} = \frac{1}{2}\left[\sum_{r=1}^{i-1} a_r^{-1} - (j-1) + \sum_{r=j}^{2} a_r \right]; 1 \le i < 3, i < j \le 3$$

where

$$a_i = \frac{1 - \sum_{r=1}^{i} p_r}{\sum_{r=1}^{i} p_r}$$

SEEPS: a multi-category score which utilises a climatology

- See Rodwell et al. (2010), Haiden et al. (2012), North et al (2013), North et al (2021).
- Multi-category score
  - **Typically 3-categories**; for precipitation use: dry, light and heavy
  - Can be written as the mean of two 2-category scores assessing the dry/light and light/heavy boundaries
  - Can be calculated at stations, or on a grid, provided there is a climatology available
  - Categories are combined using a weighting matrix.



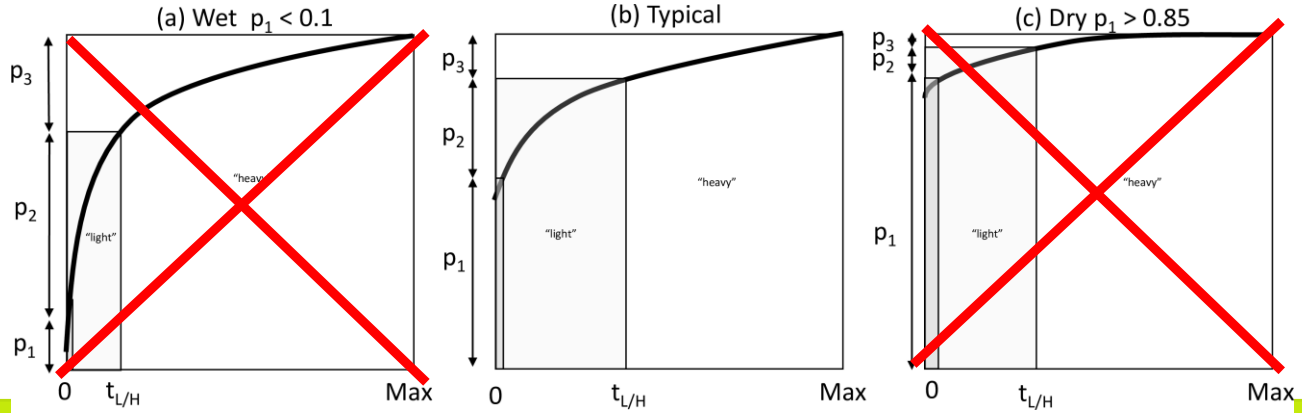
From North et al. (2021)

Commonly formulated as an error score with 3 categories: dry, light, heavy with 1-p_1 split so that $p_2 = 2 * p_3$

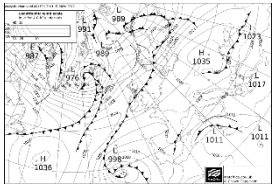
$$(s^S_{ij}) = \frac{1}{2}\begin{Bmatrix} 0 & \frac{1}{1-p_1} & \frac{1}{p_3}+\frac{1}{1-p_1} \\ \frac{1}{p_1} & 0 & \frac{1}{p_3} \\ \frac{1}{p_1}+\frac{1}{1-p_3} & \frac{1}{1-p_3} & 0 \end{Bmatrix}.$$

The concept behind SEEPS

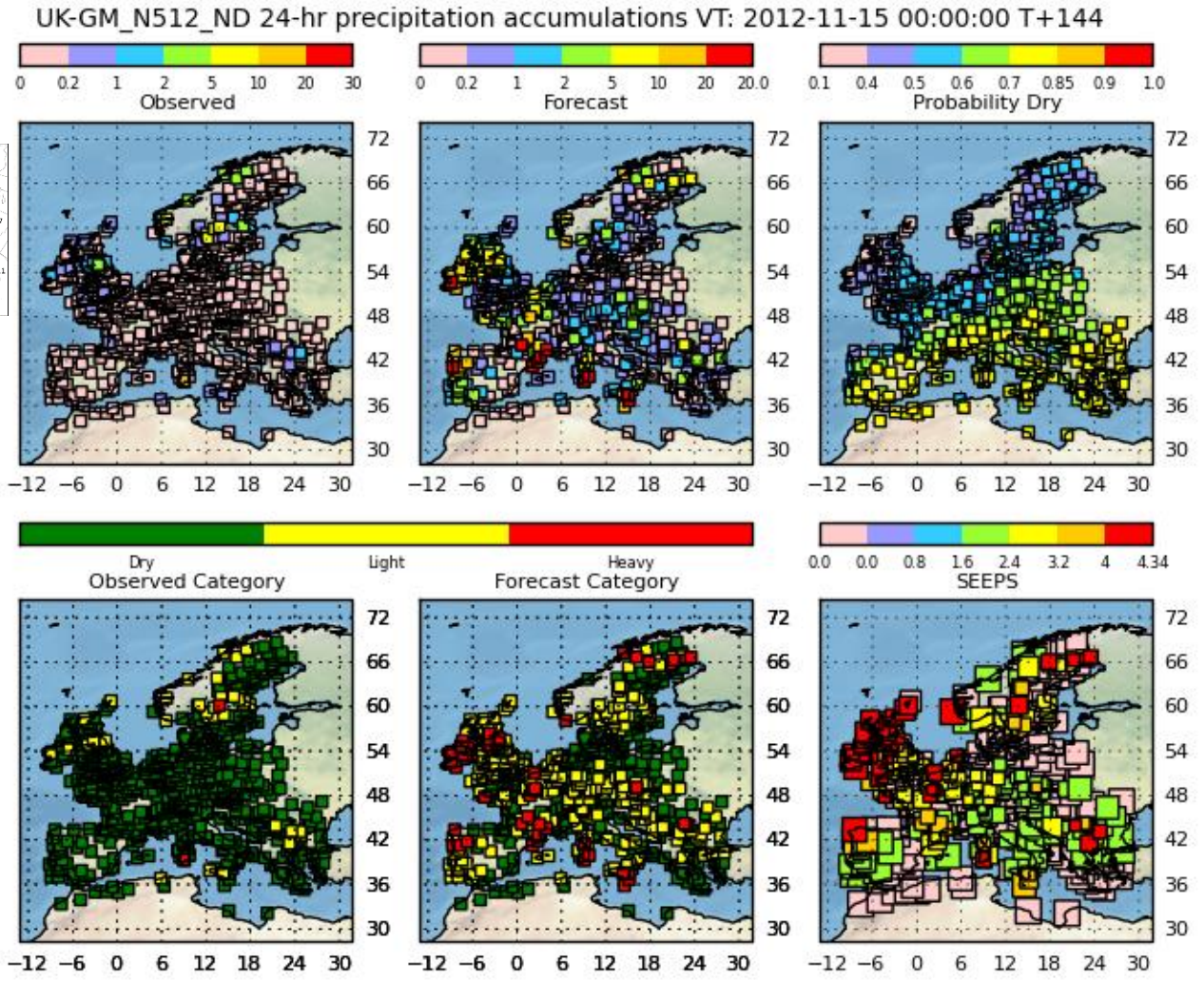
- Defining the **boundaries** between the categories is very **flexible**.
- Designed to be insensitive to sampling uncertainties
- Error score with a perfect score of 0 (usually 1-SEEPS is plotted)
- Asymptotically equitable
- Locations where the climatology is found in the upper or lower tails can lead to strange scores, and can make any aggregated score unstable and unreliable. It is best to remove these before doing any aggregation; e.g. for precip p_1 the dry/light probability threshold is set as follows: $p_1 \in [0.1, 0.85]$



SEEPS as a diagnostic tool



- The score and the components can be plotted to help diagnose systematic behaviour based on the climatological distributions.
- This makes it a powerful tool in addition to providing an overall error.



Multi-category events: single overall scores

- For **unordered** categories (e.g. precip type), Pierce skill score and Heidke skill score can be extended
- Their properties are not ideal, however (Livezey)
  - Form depends on forecast distribution
  - Weight all correct forecasts the same, which encourages conservative forecasts (forecasting the most common event)
- May be possible to tailor a score for a specific application, but considerable care needed

Multi-category events example

- Compute the scores for the following 3 category cloud forecasts.

Clouds forecast	Clouds observed			
	0 - 2	3 - 5	6 - 8	fc Σ
0 - 2	65	10	21	96
3 - 5	29	17	48	94
6 - 8	18	10	128	156
obs Σ	112	37	197	346

Calculate:
PC = ?
KSS = ?
HSS = ?

Met Office

Multi-category events example

- Compute the scores for the following 3 category cloud forecasts.

Clouds forecast	Clouds observed			
	0 - 2	3 - 5	6 - 8	fc Σ
0 - 2	65	10	21	96
3 - 5	29	17	48	94
6 - 8	18	10	128	156
obs Σ	112	37	197	346

~>

PSS= 0.26, -0.36, 0.86

No clouds (0-2)	Partly cloudy (3-5)	Cloudy (6-8)
B = 0.86 POD = 0.58 FAR = 0.32 F = 0.13 TS = 0.45	B = 2.54 POD = 0.46 FAR = 0.82 F = 0.25 TS = 0.15	B = 0.79 POD = 0.65 FAR = 0.18 F = 0.19 TS = 0.57

Calculate:
PC = ?
KSS = ?
HSS = ?

results:
PC = 0.61
KSS = 0.41
HSS = 0.37

Follow-up question/thought: Do you find the multi-category score easy to interpret?

Met Office

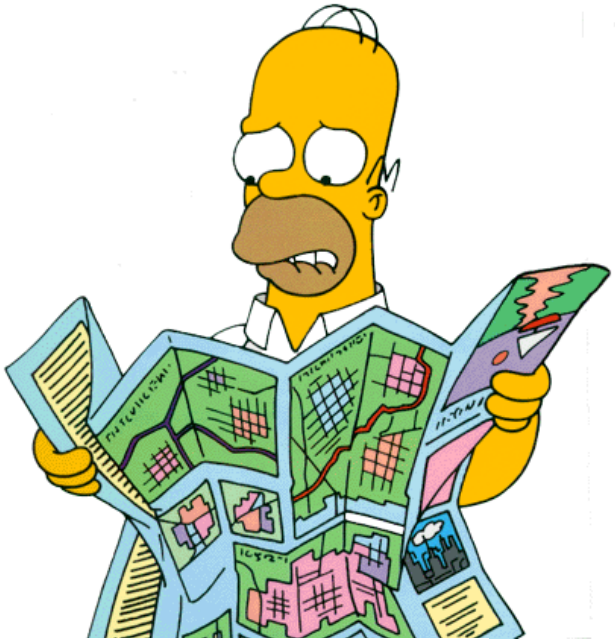
False skill? Why do we need to care?

- Introduced in the process of **aggregating in time or space** (e.g. Hamill and Juras 2006)
  - Space: mixing locations which have a very different base rate (climatology).
  - Time: mixing seasons
- This implies a **threshold dependence**: e.g.
  - *Push the threshold too high*: some regions will never contribute if values above that threshold are not climatologically possible/likely; only a few locations dominate the statistics and the overall score is not representative. (i.e. even if we are tempted to do this to measure “extremes”, real care is needed to only aggregate over locations which have similar climatologies)
  - *Keep the threshold low to be inclusive*: all places can contribute but you may not be measuring what you want, and you may still be mixing locations with different base rates.

Bottom line: aggregating in space and/or time requires careful thought to be meaningful.

Met Office

Exercise: Homer the aviation forecaster



- Springfield airport often closed due to fog
- **Homer's new job:**
  Forecast whether it will be foggy enough the next day to shut the runway
- *Six months later...*
- **Mayor Quimby asks:**
  Are Homer's forecasts any better than just tossing a coin?
  If they are, *how much* better? Is he improving?

Homer's performance

December to May

Fog forecast?	Fog observed?		
	Yes	No	Total
Yes	35	56	91
No	21	70	91
Total	56	126	182

- Base rate — $s = (35 + 21) / 182 = 0.308$
- Bias — $B = (35+56) / (35 + 21) = 1.625$
- POD — $POD = 35 / 56 = 0.625$
- POFD — $POFD = 56 / 126 = 0.444$
- Peirce skill score — $PSS = 0.625 - 0.444 =$ **0.18**

Q: does Homer have reason to celebrate?

Lisa isn’t convinced and wants a closer look...

Homer's performance

Lisa partitions the events by season….

Q: Recompute the scores. What do they tell you?

December to February

Fog forecast	Fog observed?		
	Yes	No	Total
Yes	33	43	76
No	7	7	14
Total	40	50	90

March to May

Fog forecast	Fog observed?		
	Yes	No	Total
Yes	2	13	15
No	14	63	77
Total	16	76	92

Met Office

Homer's performance

Verdict? s is very different, i.e. mixing very different climatologies. Over-forecasting in winter gives a high POD but also high POFD with POFD> POD. Spring POD and POFD much lower, though overall signal is the same → ***No*** skill in either seasons. Example where aggregating inappropriately can be detrimental to understanding performance, giving an impression of skill when there is none.

December to February

Fog forecast	Fog observed?		
	Yes	No	Total
Yes	33	43	76
No	7	7	14
Total	40	50	90

s = 40 / 90 = 0.444

B = 76 / 40 = 1.9

POD = 33 / 40 = 0.825

POFD = 43 / 50 = 0.860

PSS = 0.825 − 0.860 = −0.04

March to May

Fog forecast	Fog observed?		
	Yes	No	Total
Yes	2	13	15
No	14	63	77
Total	16	76	92

16 / 92 = 0.174

16/15 = 1.067

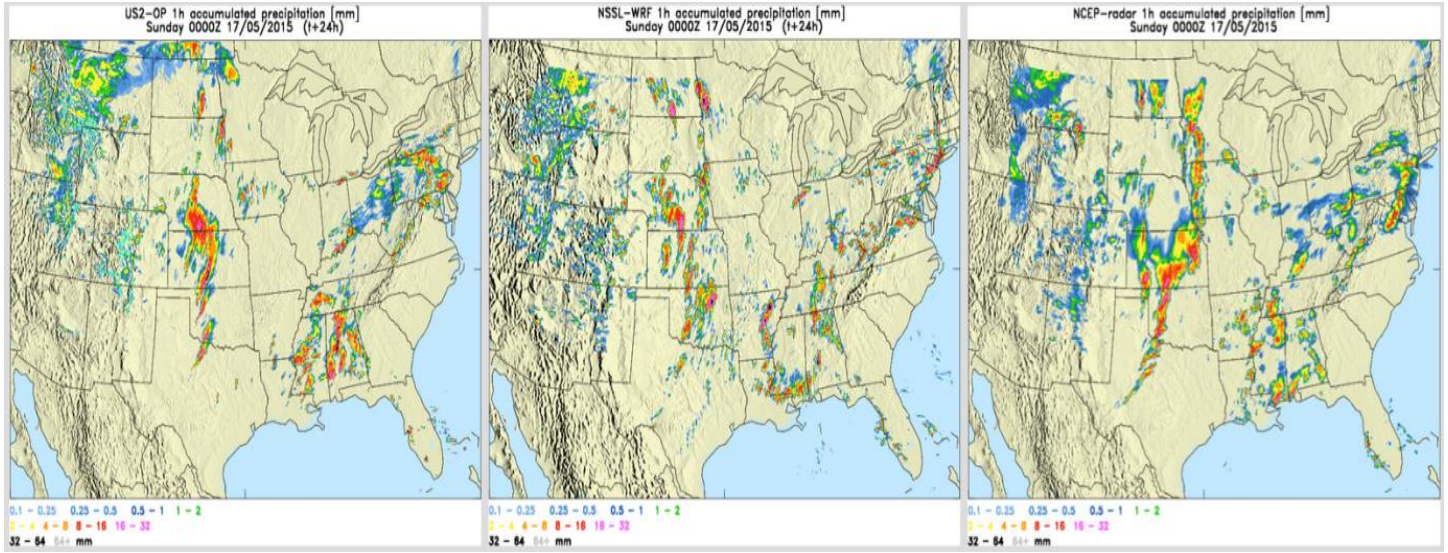
2 / 16 = 0.125

13 / 76 = 0.171

0.125 − 0.171 = −0.05

Double penalty and km-scale models

- The detail looks very realistic but is it right?

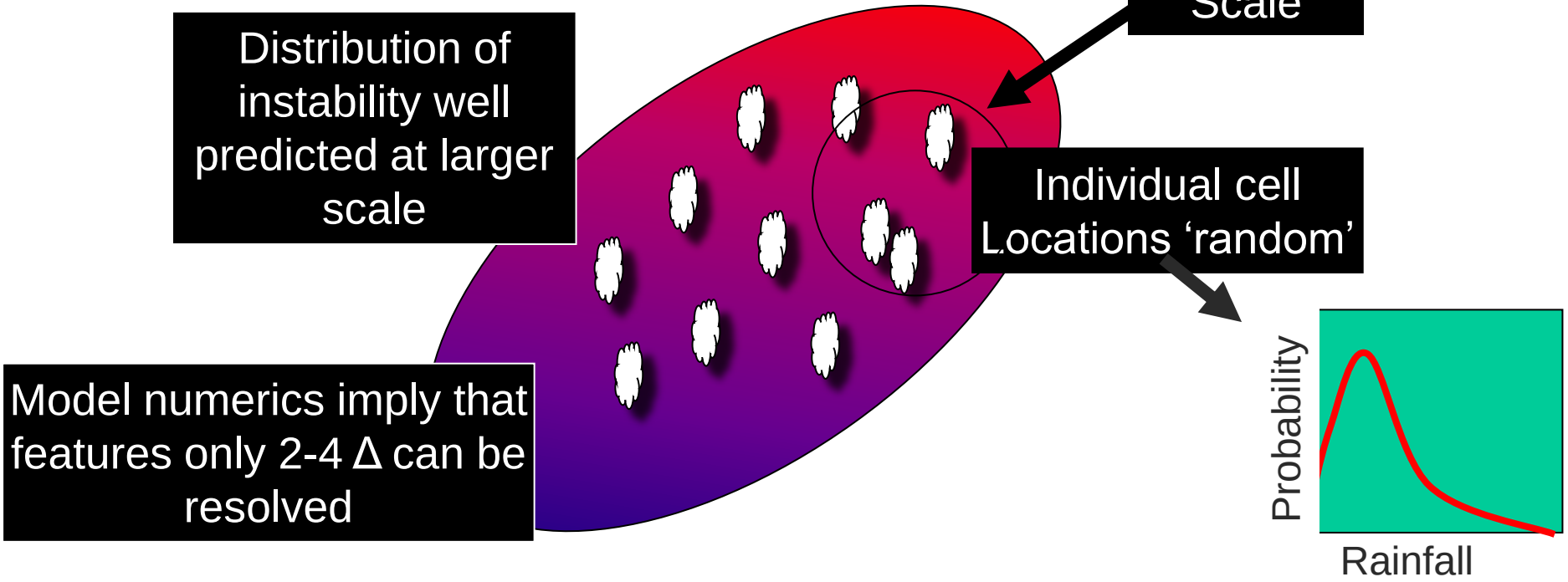


Q: What happens if you calculated categorical scores based on a pointwise comparison between these models and the radar?

From Kain et al. 2016

Met Office

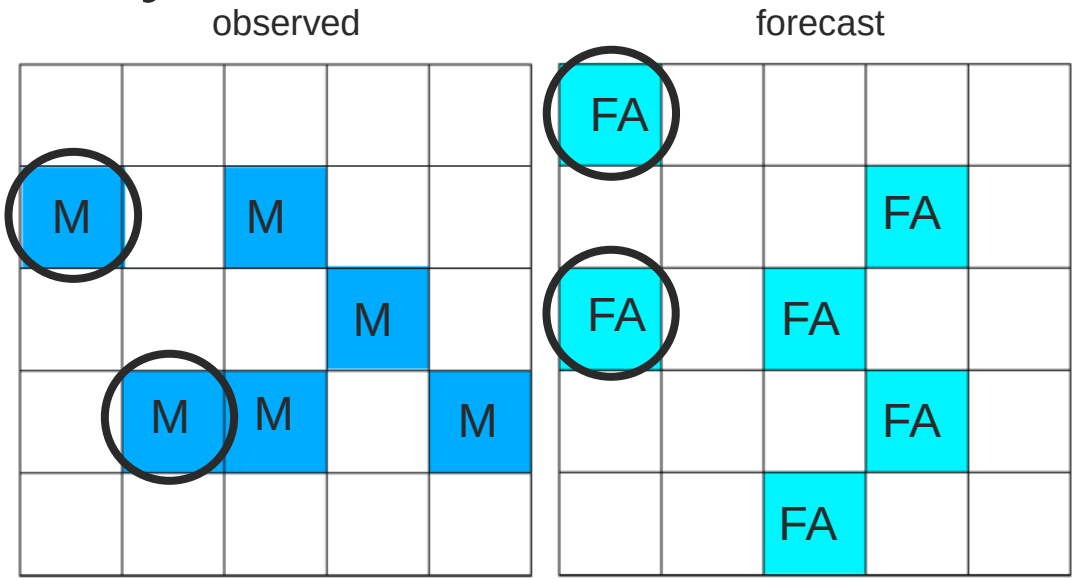
We shouldn't believe high-resolution at or near the grid scale: so, we shouldn't attempt to verify it!



Courtesy of Peter Clark

How the double penalty comes about…

- Consider this example where the number of threshold exceedances (6) are the same but none overlap.
- Each of the exceedances has a “match” which is displaced.
- In this case a pointwise comparison will **count the error 12 times (twice the threshold exceedances)** due to this displacement, i.e. the forecast is penalised for not having the event where it is observed, and for forecasting the event where it was not observed.



Threshold exceeded where squares are blue

More on how we mitigate against the double penalty in the spatial methods lecture

Met Office

Practical question:

What do you do when….

Two observation types give conflicting signals when comparing model A and B?

Different samples from the same observation give different results?

What do you conclude?
What about relative magnitude of differences?

Differences in SEEPS with bootstrap CIs shaded to show when A is better than B (green), A worse than B (purple), black not sig.

Met Office

Practical question:

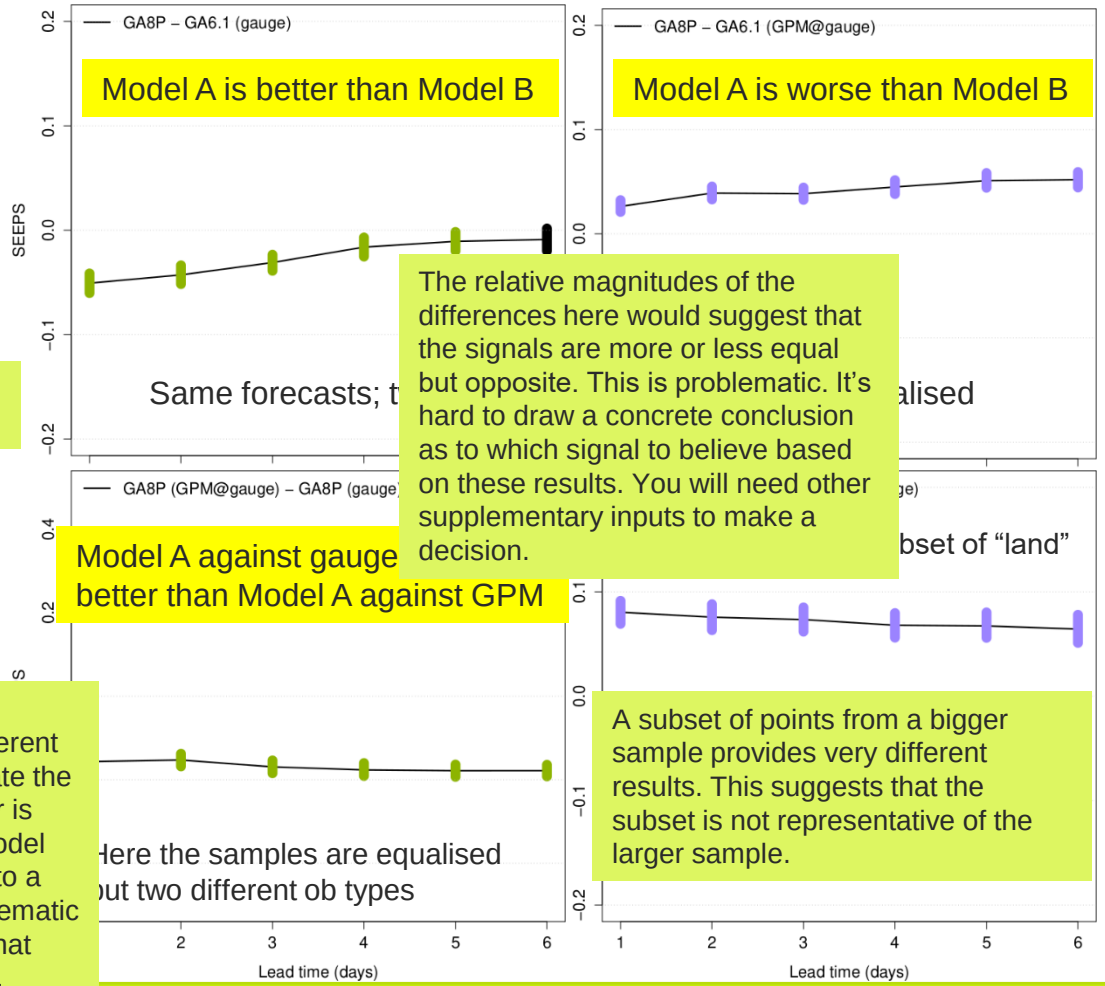
What do you do when….

This illustrates important principles of the concept of representativeness.

Two observatio
conflicting signals when comparing model A and B?

Different samples from the same observation give different results?

What do you conclu
What about relative
differences?



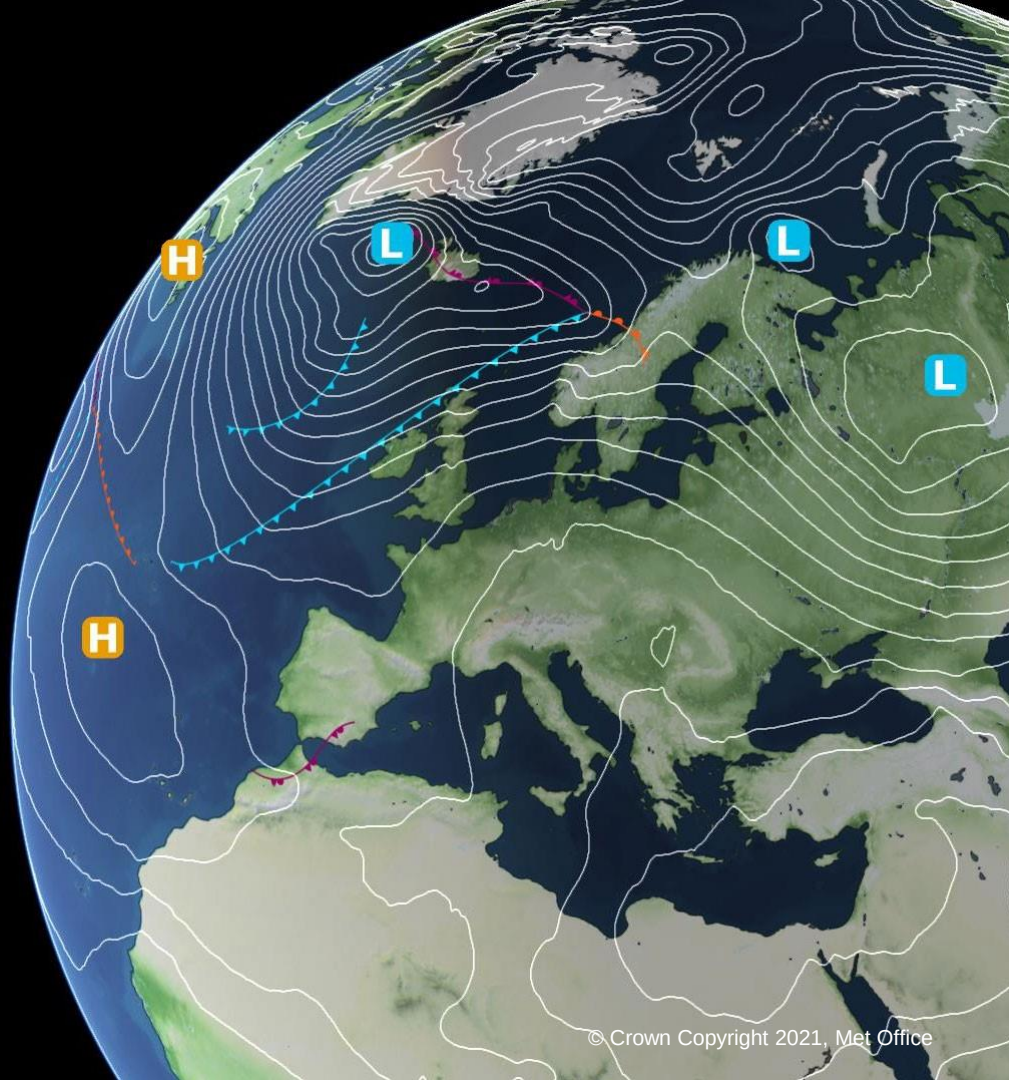
The relative magnitudes of the differences here would suggest that the signals are more or less equal but opposite. This is problematic. It's hard to draw a concrete conclusion as to which signal to believe based on these results. You will need other supplementary inputs to make a decision.

This suggests that the two observation types are very different in nature: one is a point estimate the other a grid average. The latter is closer to the spirit of what a model provides. Comparing a model to a point observation can be problematic if the point observation is not that representative of a larger area.

A subset of points from a bigger sample provides very different results. This suggests that the subset is not representative of the larger sample.

Differences in SEEPS with bootstrap CIs shaded to show when A is better than B (green), A worse than B (purple), black not sig.

Questions?



Met Office

Short score crib sheet

Basic measures/scores

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d = n

Frequency Bias Index (Bias)

$$FBI = B = \frac{(a+b)}{(a+c)}$$

- FBI > 1 over forecasting
- FBI < 1 under forecasting

Range: 0 to ∞
Perfect score = 1

Proportion Correct

$$PC = \frac{(a+d)}{n}$$

Range: 0 to 1
Perfect score = 1

- simple and intuitive
- yes and no forecasts are rewarded equally
- can be maximised by forecasting the most likely event all the time

Basic measures/scores

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d =n

Hit Rate, Probability Of Detection, Prefigurance

$$H = POD = \frac{a}{(a+c)}$$

- sensitive to misses events and hits, only
- can be improved by over forecasting
- complement score Miss Rate MS=1-H=c/(a+c)

Range: 0 to 1
Perfect score = 1

False Alarm Ratio

$$FAR = \frac{b}{(a+b)}$$

- function of false alarms and hits only
- can be improved by under forecasting

Range: 0 to 1
Perfect score = 0

Basic measures/scores

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d =n

Post agreement

$$PAG = \frac{a}{(a+b)}$$

- Complement FAR -> PAG=1-FAR
- not widely used
- sensitive to false alarms and hits

Range: 0 to 1
Perfect score = 1

False Alarm Rate, Probability of False Detection

$$F = \frac{b}{(b+d)}$$

- sensitive to false alarms and correct negative
- can be improved by under forecasting
- generally used with H (POD) to produce ROC score for probability forecasts (see later on in the week)

Range: 0 to 1
Perfect score = 0

Basic measures/scores

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d =n

Threat Score, Critical Success Index

$$TS = CSI = \frac{a}{(a+b+c)}$$

- takes into account: hits, misses and false alarms
- correct negative forecast not considered
- sensitive to climatological frequency of event

Range: 0 to 1
Perfect score = 1
No skill level = 0

Equitable Threat Score, Gilbert Skill Score (GSS)

$$ETS = \frac{(a - a_r)}{(a+b+c-a_r)}$$

$$a_r = \frac{(a+b)(a+c)}{n}$$

- it is the TS which includes the hits due to the random forecast

Range: -1/3 to 1
Perfect score = 1
No skill level = 0

Basic measures/scores

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d =n

Hanssen & Kuipper's Skill Score, True Skill Statistic (TSS), Pierce's Skill Score

$$KSS = TSS = H - F = \frac{(ad - bc)}{[(a+c)(b+d)]}$$

- popular combination of H and F
- Measures the ability to separate yes (H)and no (F) cases
- For rare events d is very large -> F small and KSS (TSS) close to POD (H)
- Related to ROC (Relative Operating Characteristic)

Range: -1 to 1
Perfect score = 1
No skill level = 0

Heidke Skill Score

$$HSS = \frac{2(ad - bc)}{[(a+c)(c+d) + (a+b)(b+d)]}$$

- Measures fractional improvements over random chance
- Usually used to score multi-category events

Range: - ∞ to 1
Perfect score = 1
No skill level = 0

Basic measures/scores

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d =n

Odds Ratio

$$OR = \frac{ad}{bc}$$

- measures the forecast probability(odds) to score a hit (H) compared to giving a false alarm (F)

$$OR = \frac{\left[\frac{H}{1-H}\right]}{\left[\frac{F}{1-F}\right]}$$

- independent of biases
- unbound

Range: 0 to ∞
Perfect score = ∞
No skill level = 1

Odds Ratio Skill Score

$$ORSS = \frac{(ad - bc)}{(ad + bc)} = \frac{OR - 1}{OR + 1}$$

- produces typically very high absolute skill values (because of its definition)
- Not widely used in meteorology

Range: -1 to 1
Perfect score = 1

Met Office

Practical problem:

Compute the scores for these
Two contingency tables and discuss.

Gale forecast	Gale observed		
	Yes	No	fc Σ
Yes	15	2	17
No	11	123	134
obs Σ	26	125	151

Tornado forecast	Tornado observed		
	Yes	No	fc Σ
Yes	30	70	100
No	20	2680	2700
obs Σ	50	2750	2800

Left panel: Contingency table for five months of categorical warnings against gale-force winds (wind speed > 14m/s)
Right panel: Tornado verification statistics

	GALE	TORNADO
B = (a+b)/(a+c)		
PC = (a+d)/n		
POD = a/(a+c)		
FAR = b/(a+b)		
PAG = a/(a+b)		
F = b/(b+d)		
KSS = POD-POFD		
TS = a/(a+b+c)		
ETS = $(a-a_r)/(a+b+c-a_r)$		
HSS = 2(ad-bc)/[(a+c)(c+d)+(a+b)(b+d)]		
OR = ad/bc		
ORSS = (OR-1)/(OR+1)		

Practical: results

Gale forecast	Gale observed		
	Yes	No	fc Σ
Yes	15	2	17
No	11	123	134
obs Σ	26	125	151

Tornado forecast	Tornado observed		
	Yes	No	fc Σ
Yes	30	70	100
No	20	2680	2700
obs Σ	50	2750	2800

Left panel: Contingency table for five months of categorical warnings against gale-force winds (wind speed > 14m/s)
Right panel: Tornado verification statistics

	GALE	TORNADO
B = (a+b)/()a+c)	_0.65__	_2.00_
PC = (a+d)/n	_0.91__	_0.97_
POD = a/(a+c)	_0.58__	_0.60_
FAR = b/(a+b)	_0.12__	_0.70_
PAG = a/(a+b)	_0.88__	_0.30_
F = b/(b+d)	_0.02__	_0.03_
KSS = POD-POFD	_0.56__	_0.57_
TS = a/(a+b+c)	_0.54__	_0.25_
ETS = (a-a)/(a+b+c-a)	_0.48__	_0.24_
HSS = 2(ad-bc)/[(a+c)(c+d)+(a+b)(b+d)]	_0.65__	_0.39_
OR = ad/bc	_83.86_	_57.43_
ORSS = (OR-1)/(OR+1)	_0.98__	_0.97_