A sequential decomposition of proper scoring rules

SAM ALLEN

CHRIS FERRO

FRANK KWASNIOK

UNIVERSITY OF EXETER

MeteoSwiss

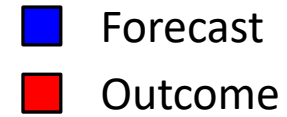
Forecasting

- The process of making predictions based on the available information

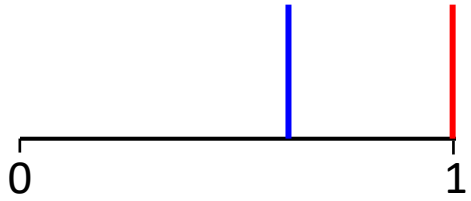
- Forecasts can be made for a range of different kinds of outcomes

- For example:
  - Will it rain on Saturday?
  - What will the global population be in 2050?
  - How many major football championships will England win this century?

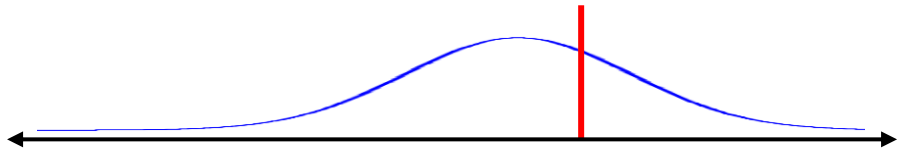
Forecasting



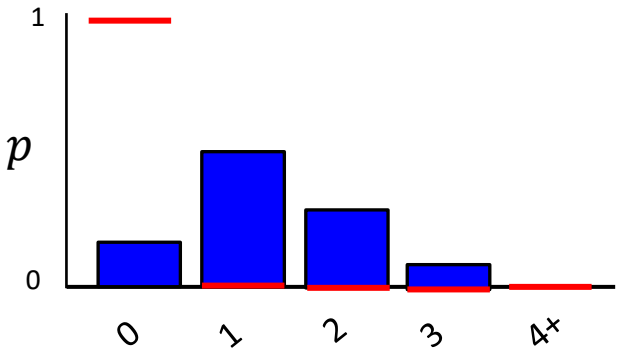
- Will it rain on Saturday?



- What will the global population be in 2050?



- How many major football championships will England win this century?



Forecast Verification

- How do we assess the performance of a forecaster?
- We need to consider several aspects of the forecaster's performance
- It is convenient to objectively quantify the accuracy of a forecaster
- This allows us to rank and compare competing forecasters

Scoring Rules

- A scoring rule is a function that maps from the joint space of the forecasts $\mathcal{F}$ and the outcomes Ω to the real line

$$S: \mathcal{F} \times \Omega \to \mathbb{R} \cup \{-\infty, \infty\}$$

- Scoring rules therefore provide a single numerical measure of the forecast's performance
  - We consider negatively oriented scoring rules (i.e. lower score = better forecast)

- For example: $S(x, y) = (x - y)^2$ $x, y \in \mathbb{R}$

Scoring Rules

- A scoring rule is a function that maps from the joint space of the forecasts $\mathcal{F}$ and the outcomes Ω to the real line

$$S: \mathcal{F} \times \Omega \to \mathbb{R} \cup \{-\infty, \infty\}$$

- Scoring rules therefore provide a single numerical measure of the forecast's performance
  - We consider negatively oriented scoring rules (i.e. lower score = better forecast)

- For example:

$S(x, y) = (x - y)^2$	$x, y \in \mathbb{R}$	A 'good' scoring rule
$S(x, y) = x + y$	$x, y \in \mathbb{R}$	A 'bad' scoring rule

Proper Scoring Rules

- Just as forecasts should exhibit desirable properties, so too should the scoring rules

- The scoring rule should be optimised by a forecast that is equivalent to the true process governing the outcomes

$$E_{Y\sim G}[S(G,Y)] \leq E_{Y\sim G}[S(F,Y)]$$

- This encourages a forecaster to issue their true beliefs

- Such scoring rules are termed proper

Decompositions of Scoring Rules

- Scoring rules provide one overall measure of the forecast accuracy
- The value of a forecast will change depending on how it is to be used
- Decompositions of scoring rules exist that partition the expected score into different terms…
- …each of which assesses a distinct aspect of the forecaster's performance

Decompositions of Scoring Rules

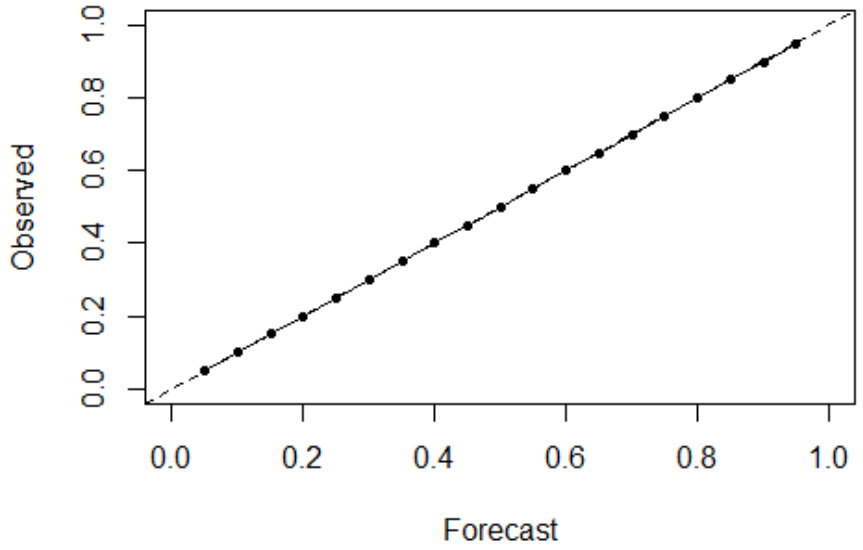
- The most well-known decomposition is into uncertainty, resolution and reliability terms:

$$E[S(F,Y)] = UNC_Y - RES_F + REL_F$$ (Bröcker, 2009)

- UNC_Y is the uncertainty - the inherent uncertainty in the forecast setting
- RES_F is the resolution - the information content of the forecaster
- REL_F is the reliability - the (mis-)calibration of the forecaster
  - The information lost due to miscalibration
- For proper scoring rules, all of these terms are non-negative

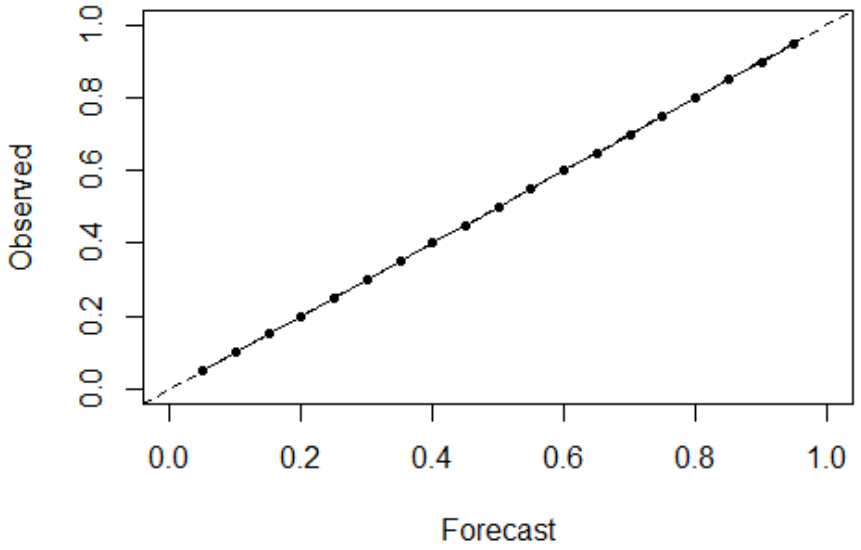
Reliability

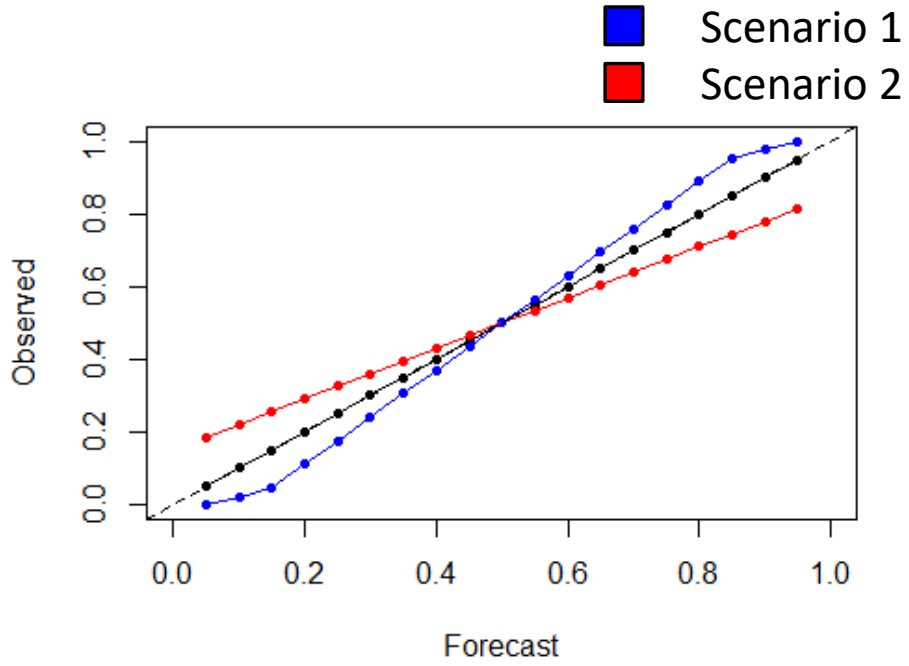
- What does it mean for a forecast to be reliable?
  - If an event is forecast with probability p, then it should occur with probability p
- Reliability diagrams allow us to visualise the calibration of probability forecasts



Reliability

- But even if a forecaster is reliable overall, it may exhibit conditional biases





Conditional Verification

- Forecast performance may change under different circumstances
- It would be beneficial if the forecaster (and the forecast user) knew in what circumstances their forecasts performed poorly
- Consider a set of mutually exclusive, collectively exhaustive scenarios $A \in \{A_1, ..., A_J\}$
- These scenarios could be seasons, weather regimes, spatial regions etc.

Conditional Verification

- Forecasters could calculate the terms of the decomposition separately in each scenario

$$E\left[S(F,Y)|A_j\right] = UNC_{Y|A_j} - RES_{F|A_j} + REL_{F|A_j}$$

Conditional Verification

- Forecasters could calculate the terms of the decomposition separately in each scenario

$$E\left[S(F,Y)|A_j\right] = UNC_{Y|A_j} - RES_{F|A_j} + REL_{F|A_j}$$

- Then take the expectation of each term across the different scenarios to recover the expected score

$$E[S(F,Y)] = UNC_{Y|A} - RES_{F|A} + REL_{F|A}$$

Conditional Verification

- Forecasters could calculate the terms of the decomposition separately in each scenario

$$E\left[S(F,Y)|A_j\right] = UNC_{Y|A_j} - RES_{F|A_j} + REL_{F|A_j}$$

- Then take the expectation of each term across the different scenarios to recover the expected score

$$E[S(F,Y)] = UNC_{Y|A} - RES_{F|A} + REL_{F|A}$$

- $UNC_{Y|A}$ is the conditional uncertainty given the chosen scenarios
- $RES_{F|A}$ is the conditional resolution
- $REL_{F|A}$ is the conditional reliability

Conditional Verification

- However this is not the same as the previous decomposition

$$E[S(F,Y)] = UNC_Y - RES_F + REL_F$$

$$E[S(F,Y)] = UNC_{Y|A} - RES_{F|A} + REL_{F|A}$$

- For example, a forecast may be unconditionally calibrated: $REL_F = 0$

  but still exhibit conditional biases: $REL_{F|A} > 0$
  - $REL_{F|A}$ is a stronger criterion for calibration than REL_F

- We have two different decompositions of proper scoring rules into uncertainty, resolution and reliability
  - Which should we use?
  - How do they relate to each other?

A New Decomposition

- The two decompositions can be combined to form a new decomposition

$$E[S(F,Y)] = UNC_{Y|A} - RES_{F|A} + REL_{F|A}$$

A New Decomposition

- The two decompositions can be combined to form a new decomposition

$$E[S(F,Y)] = UNC_{Y|A} + RES_A - RES_A - RES_{F|A} + RES_{A|F} + REL_{F|A} - RES_{A|F}$$

- RES_A is the amount of information provided by the different scenarios
- $RES_{A|F}$ is the amount of information provided by the different scenarios…

… that is **not** already captured by the forecast

A New Decomposition

- The two decompositions can be combined to form a new decomposition

$$E[S(F,Y)] = \underbrace{\{UNC_{Y|A} + RES_A\}}_{UNC_Y} - \underbrace{\{RES_A + RES_{F|A} - RES_{A|F}\}}_{RES_F} + \underbrace{\{REL_{F|A} - RES_{A|F}\}}_{REL_F}$$

A New Decomposition

- The two decompositions can be combined to form a new decomposition

$$E[S(F,Y)] = \underbrace{\{UNC_{Y|A} + RES_A\}}_{UNC_Y} - \underbrace{\{RES_A + RES_{F|A} - RES_{A|F}\}}_{RES_F} + \underbrace{\{REL_{F|A} - RES_{A|F}\}}_{REL_F}$$

- Two terms have been both added and subtracted from the conditional decomposition
- This decomposition has five terms instead of three
- These do not contribute to the overall score, but are informative nonetheless

A New Decomposition

$$E[S(F,Y)] = \underbrace{\{UNC_{Y|A} + RES_A\}}_{UNC_Y} - \underbrace{\{RES_A + RES_{F|A} - RES_{A|F}\}}_{RES_F} + \underbrace{\{REL_{F|A} - RES_{A|F}\}}_{REL_F}$$

- The uncertainty UNC_Y has been factorised into within- and between-scenario components
- The forecast resolution RES_F is now the sum of
  1. The information provided by the scenarios
  2. The additional information provided by the forecasts
  3. Minus the information provided by the scenarios that is **not** captured by the forecast
- The overall reliability REL_F has been decomposed into two terms
  - If both terms are zero, then this increases the information content of the forecast

A Sequential Decomposition

- The decomposition can be extended easily to condition on more than one set of scenarios

- Consider a second set of scenarios $B \in \{B_1, \ldots, B_L\}$

$$E[S(F,Y)] = \{UNC_{Y|A,B} + RES_{A,B}\} - \{RES_{A,B} + RES_{F|A,B} - RES_{A,B|F}\} + \{REL_{F|A,B} - RES_{A,B|F}\}$$

A Sequential Decomposition

- The decomposition can be extended easily to condition on more than one set of scenarios

- Consider a second set of scenarios $B \in \{B_1, \ldots, B_L\}$

$$E[S(F,Y)] = \{UNC_{Y|A,B} + RES_{A,B}\} - \{RES_{A,B} + RES_{F|A,B} - RES_{A,B|F}\} + \{REL_{F|A,B} - RES_{A,B|F}\}$$

- The information provided by A and B can itself be decomposed into contributions from each

$$RES_{A,B} = RES_A + RES_{B|A} = RES_B + RES_{A|B}$$

- This is also true for the conditional resolution term $RES_{A,B|F}$

Analysis of Information

- We can use this to sequentially analyse the sources of information in the forecast
- This is similar to the Analysis of Variance (ANOVA) framework
  - Used to quantify the amount of information provided by different factors in linear regression models
- The out-of-sample ANOVA is equivalent to the decomposition of the squared error into uncertainty, resolution and reliability terms

Analysis of Information

- Decompositions of scoring rules can be thought of as a generalised analysis of variance framework
- The average score represents the amount of information that is not explained by the forecast
- The classical decomposition terms yield a measure analogous to the R^2 goodness of fit statistic
  - Equivalent to the skill score relative to an unconditional forecaster

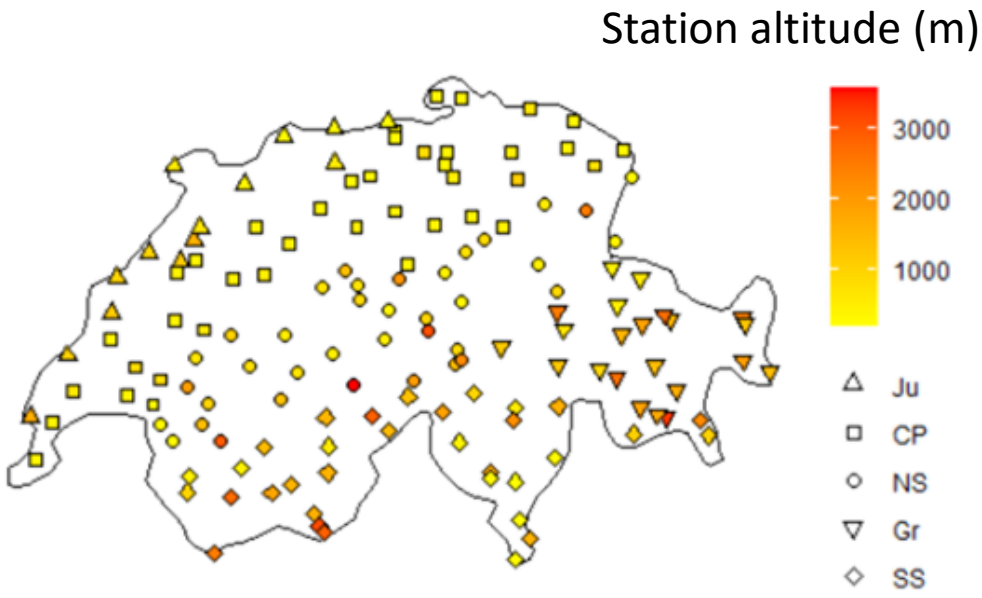
$$R^2 \approx \frac{RES_F - REL_F}{UNC_Y}$$

- The terms of the new decomposition can be used to define an analogue to the partial R^2

$$Partial\ R^2 \approx \frac{RES_{A|F}}{UNC_Y - RES_F}$$

Case Study

- We present an example using MeteoSwiss temperature exceedance forecasts
- Consider forecasts at a lead time of 3 days
- In the summer seasons (JJA) of 2016-2020
- At 146 weather stations over Switzerland



$n_{k\blacksquare}$ = Number of times P_k was forecast

$\bar{y}_{k\blacksquare}$ = Mean outcome given P_k was forecast

Case Study

- We use the Brier score (Brier, 1950) to assess these probabilistic forecasts:

$$BS(f, y) = (f - y)^2, \qquad f \in [0, 1],\ y \in \{0, 1\}$$

- Consider $f \in \{P_1, \ldots, P_K\}$, then the average Brier score over n forecast cases can be written as

$$\frac{1}{n}\sum_{i=1}^{n} BS(f_i, y_i) = \underbrace{\bar{y}(1-\bar{y})}_{UNC_Y} - \underbrace{\sum_{k=1}^{K} \frac{n_{k\bullet}}{n}(\bar{y}_{k\bullet} - \bar{y})^2}_{RES_F} + \underbrace{\sum_{k=1}^{K} \frac{n_{k\bullet}}{n}(P_k - \bar{y}_{k\bullet})^2}_{REL_F}$$

(Murphy, 1973)

Case Study

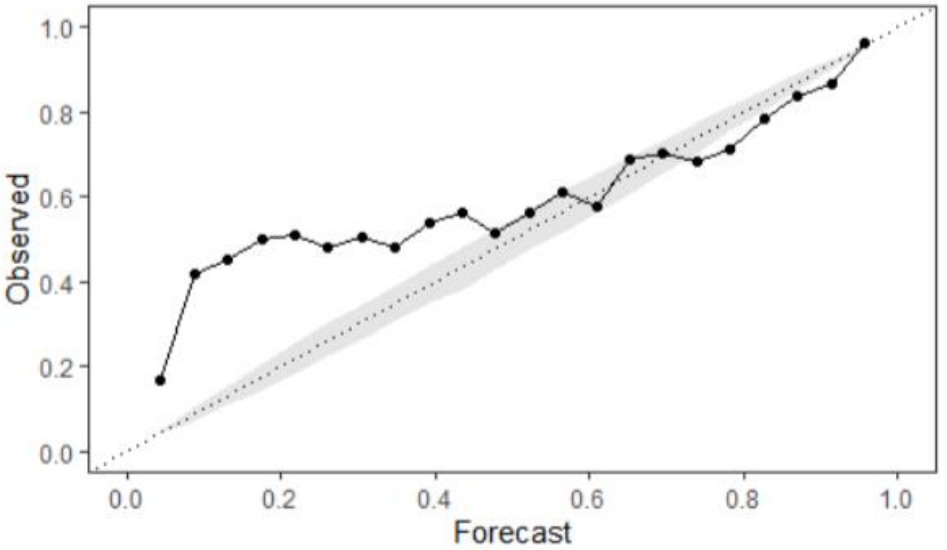
- Consider the performance of MeteoSwiss' COSMO-2E ensemble prediction system

- Ensemble forecasts are a collection of point forecasts

- We take the forecast probability to be the proportion of ensemble members that exceed the threshold of interest
  - These are commonly used to construct heat warnings

- The reliability diagram indicates that the prediction system is not reliable

Temp. threshold = 20°C

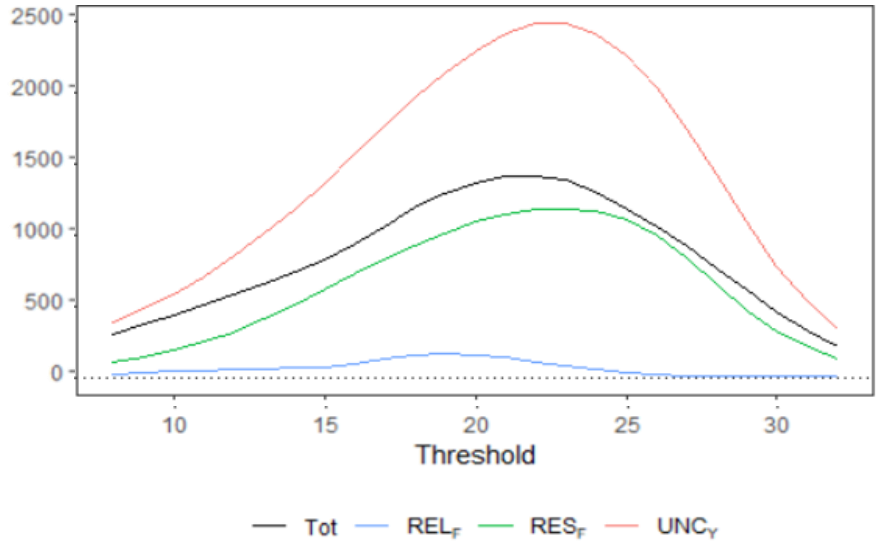


Case Study

- The classical decomposition of the Brier score reinforces this

	UNC_Y	RES_F	REL_F
COSMO	2297.0	1090.7	141.8

These terms have been scaled by 10^4 to aid interpretation



We can see how these terms change as we alter the threshold of interest

Case Study

- The classical decomposition of the Brier score reinforces this

	UNC_Y	RES_F	REL_F
COSMO	2297.0	1090.7	141.8

These terms have been scaled by 10^4 to aid interpretation

- The forecasts explain $\frac{RES_F - REL_F}{UNC_Y} \approx 41.3\%$ of the uncertainty in the outcome

- A (perfect) re-calibration scheme would increase this to $\frac{RES_F}{UNC_Y} \approx 47.5\%$

Case Study

- We re-calibrate the forecasts using a simple logistic regression

	UNC_Y	RES_F	REL_F
COSMO	2297.0	1090.7	141.8
Re-calibrated	2297.0	1082.1	63.8

- The resulting forecasts are more reliable, though the simple post-processing approach is not perfect

- The information in the COSMO forecast has been retained

Case Study

- Consider the decomposition conditional on the altitude of the weather station

	UNC_Y	RES_F	REL_F
COSMO	2297.0	1090.7	141.8
Recalibrated	2297.0	1082.1	63.8

$UNC_{Y\|A}$	RES_A	$RES_{F\|A}$	$RES_{A\|F}$	$REL_{F\|A}$
1411.0	886.0	378.9	174.2	316.0
1411.0	886.0	372.5	176.5	240.2

$Total$
1348.1
1278.7

- The altitudes are grouped into the following categories:
  1. <500m
  2. [500, 1000)
  3. [1000, 1500)
  4. [1500, 2000)
  5. ≥2000m

Case Study

	Uncertainty UNC_Y	Resolution RES_F	Reliability REL_F	Conditional Uncertainty $UNC_{Y\|A}$	Scenario Information RES_A	Conditional Forecast Information $RES_{F\|A}$	Conditional Scenario Information $RES_{A\|F}$	Conditional Reliability $REL_{F\|A}$	Brier Score $Total$
COSMO	2297.0	1090.7	141.8	1411.0	886.0	378.9	174.2	316.0	1348.1
Recalibrated	2297.0	1082.1	63.8	1411.0	886.0	372.5	176.5	240.2	1278.7

- The altitude groups explain $\frac{RES_A}{UNC_Y} \approx 38.6\%$ of the variation in the temperature exceedance

Case Study

	UNC_Y	RES_F	REL_F
COSMO	2297.0	1090.7	141.8
Recalibrated	2297.0	1082.1	63.8

$UNC_{Y\|A}$	RES_A	$RES_{F\|A}$	$RES_{A\|F}$	$REL_{F\|A}$	$Total$
1411.0	886.0	378.9	174.2	316.0	1348.1
1411.0	886.0	372.5	176.5	240.2	1278.7

- The altitude groups explain $\frac{RES_A}{UNC_Y} \approx 38.6\%$ of the variation in the temperature exceedance
- $\frac{RES_{A|F}}{RES_A} \approx 19.9\%$ of this variation is **not** explained by the forecasts

Case Study

	UNC_Y	RES_F	REL_F
COSMO	2297.0	1090.7	141.8
Recalibrated	2297.0	1082.1	63.8

$UNC_{Y\|A}$	RES_A	$RES_{F\|A}$	$RES_{A\|F}$	$REL_{F\|A}$	$Total$
1411.0	886.0	378.9	174.2	316.0	1348.1
1411.0	886.0	372.5	176.5	240.2	1278.7

- Although the reliability is increased, the re-calibrated forecasts are still unreliable conditional on the altitude groups

Case Study

	UNC_Y	RES_F	REL_F	$UNC_{Y\|A}$	RES_A	$RES_{F\|A}$	$RES_{A\|F}$	$REL_{F\|A}$	$Total$
COSMO	2297.0	1090.7	141.8	1411.0	886.0	378.9	174.2	316.0	1348.1
Recalibrated	2297.0	1082.1	63.8	1411.0	886.0	372.5	176.5	240.2	1278.7

- Although the reliability is increased, the re-calibrated forecasts are still unreliable conditional on the altitude groups
- A recalibration scheme that incorporated altitude information would explain a further

$$\frac{RES_{A|F}}{UNC_Y - RES_F} \approx 14.5\%$$ of the uncertainty in the outcome

Case Study

- We apply a second logistic regression that uses the altitude as an additional predictor

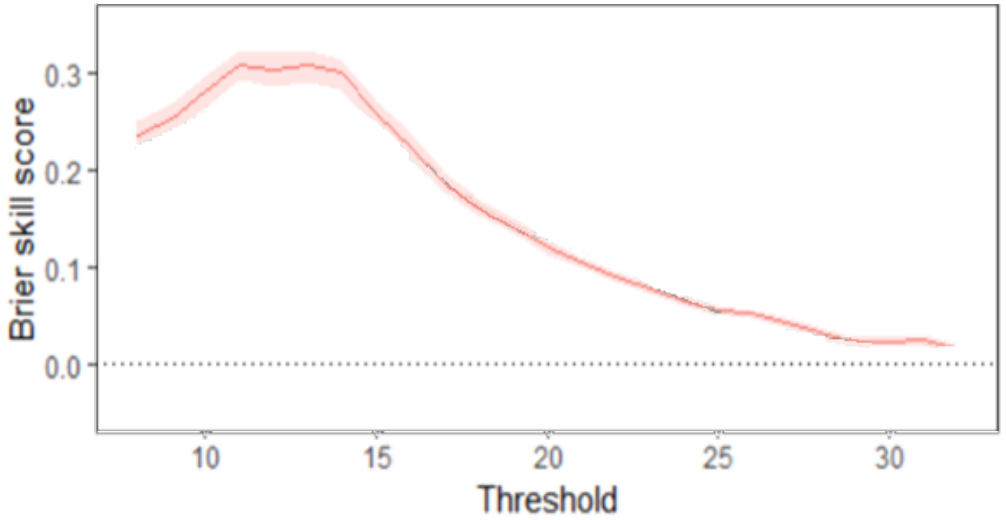
	UNC_Y	RES_F	REL_F
COSMO	2297.0	1090.7	141.8
Recalibrated	2297.0	1082.1	63.8
ADPP	2297.0	1238.8	67.9

$UNC_{Y\|A}$	RES_A	$RES_{F\|A}$	$RES_{A\|F}$	$REL_{F\|A}$	$Total$
1411.0	886.0	378.9	174.2	316.0	1348.1
1411.0	886.0	372.5	176.5	240.2	1278.7
1411.0	886.0	369.9	17.2	85.1	1126.2

- The altitude-dependent post-processing (ADPP) model explains $1 - \frac{RES_{A|F}}{RES_A} \approx 98.1\%$ of the variation due to the altitude groups
- This reduction in conditional biases manifests as an increase in resolution (and a decrease in score)

Case Study

- Consider the improvement gained by including altitude information into post-processing



- There is benefit at all thresholds considered here, particularly at low temperatures (up to 30% better)
- The altitude groups explain more about when the temperature will be low, than when it will be high

Conclusions

- We introduce a new decomposition of scoring rules that combines two possible approaches

- It allows the forecaster to quantify the information provided by different sources

  and to identify the presence of conditional biases, while relating this back to the overall reliability

- We propose that decompositions of scoring rules generalise the analysis of variance framework

- The identification of conditional biases can be used to improve statistical post-processing models
- We illustrate this using a case study on MeteoSwiss temperature exceedance forecasts

References

- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1-3.
- Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *135*(643), 1512-1519.
- Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, *12*(4), 595-600.