15:00UTC Session on Metaverification

Harold Brooks: The Relationship Between ROC, Performance, and the Quality-Decision Threshold Diagrams

Dominique Brunet:

If I got it right, the quality of the forecast has not improved in the last 20 years (on the warning side)?

HB: (?)Yes. We are looking to draw additional information from probabilistic considerations.

DB: Slide 'modelling the problem' on d': Is there a Gaussian assumption? Is that good?

Good in some ways, bad in others. There are many possible ways, including a Logistic assumption. This idea is used in many fields, and many different thresholds are important. The Gaussian assumption facilitates analyses.

Barbara Casati:

Plot dots for periods: Interpretation? It seems in the red, POD is really high, then you regress in 2012-2019?

HB:

Two things happened. In 2012 the central region, in 2013 southern region, raised the threshold needed to put a warning out. That reduced false alarms.

Other effects: Radar updates started to become more frequent. As a result, forecasters waited longer to issue warnings with new information coming in frequently.

The default warning length changed from 30 to 40 minutes, with effects on the probability of detection and false alarm rates.

Raghavendra Ashrit:

Progress from grey to green to blue and so on in zROC plot – how can we inpret crossing the diagonal?

HB:

As we move to the upper left, that's good. It looks like we started training forecasters better from 1990.

Other effects/improvements: Over time there was less manual interaction, so more warnings were issued. Also, a lot more storms did not make tornados than we thought.

Tom Robinson:

Increase of emphasis to reduce false alarms in 2012/2013: What did your clients think of the change?

HB:

Change of emphasis was spurred by the tornado in Joplin, MI, 2011. There were complaints about too many alarms, but that was a rare forecasting event.

Before, from 2008-2011, there were 24 consecutive warnings with no tornado (from previously 1 in a long time). There are false alarm problems in limited regions.

People sensitive to false alarms were delighted, others were not. Of course, the probability of detection and false alarm rates changed.

Carlo Cafaro (16:41) :

Hi Harold, thanks for the great talk. Could be also because there have been more reports of tornado with time? (and so fewer false alarms?) thanks.

HB (16:47):

Carlo – in short, not for the recent change. Our tornado count hasn't changed much over the last 15 years. It is actually a little lower since 2012.

Kenric Nelson: Detecting over-confidence in weather forecasts

Dominique Brunet:

Why is the generalized mean used as a measure for robustness?

KN:

The metrics come from the generalization of the logarithmic function. There is a translation between the negative and positive values. The cited theoretical papers give a derivation.

Yawei Ning: A new skill score for quantifying the uncertainty in multi-category precipitation forecasts

Dominique Brunet (17:11):

What is Li?

YN (17:45):

Li is the category index. For example, L1 represents No rain. L2 represents Light rain. Sorry for misunderstanding.

Raghavendra Ahrit:

Is the frequency bias also removed from the four different models?

-

Kenric Nelson:

I am puzzled why the metric NMI is going up with longer lead time? Could you clarify the metric.

-

KN:

Later you show the trend going in the other direction. Do you know why? Intuition on L3, why is the trend the other way?

-

Dominique Brunet (17:23):

My experience with mutual information (for image processing applications) is that it is sometimes hard to estimate the joint probability distribution of X and Y. How do you estimate it in your case and is the estimations pesents any issue at all?

YN (17:38):

We use the mutual information of discrete forms, and needn't to estimate the parameters. (The probability P(x, y) is measure by the ratio of their occurrence in). I do not know if it help.

DB (17:40):

Yes, it answers the question. I understand you are binning the rain amount into discrete categories, so as long the rarest category is sufficiently sampled the MI will work.

Michael Sharpe: A complementary measure to assess temporal uncertainty within Terminal Aerodrome Forecasts

Barbara Casati (17:41):

Allowing larger and larger temporal uncertainty (a long TEMPO) is similar to allowing larger and larger neighbourhoods (though in this case they are not in space, but in time).

Harold Brooks

Did you have an opportunity to talk to airline dispatchers? Different ways to write the TAFs? Is there a difference between long-haul / short-haul?

MS:

I have been working with Andre Lanyon, he worked as a forecaster for a long time. Unfortunately, we did not have the opportunity for that comparison.

Günter Mahringer:

I appreciate your work. Shortly explain, how do you apply this for BECOMING groups?

MS:

There wasn't time in the talk, unfortunately. I can share that with you separately. You can derive that in the same way.

GM:

Tempo group: You stated that a TEMPO group has no probability uncertainty. At the end, you use an assignment of a probability. But forecasters don't want to express probability. Is that adequate?

MS: It is telling you the probability that any single observation is in the alternative group for this individual TEMPO. TEMPO is not a probability. If the forecast was correct, in the user perspective: What is the probability that this TEMPO is in the alternative category, given … (comment: lost track)

This is a way of assessing uncertainty. We want to make sure that uninformative TAFs are issued less often.

BC (17:48):

Why TAFs – which pertains to continuous variables – are verified with a categorical approach? Why not visualize them with a curve (such as your diamonds) and then measure the timing error? (rather than dealing with these different lengths TEMPOs)

BC:

Answered in the discussion with GM.

Nelson Shum:

What would be your strategy to aggregate the score?

MS:

What we currently do is that we look at each station and their corresponding base rates. It is not ideal to combine stations, as that would smooth out base rates. We calculate a score over a time period for each airport independently, and then report the median airport and the main airports.

Ashrit, Raghavendra (17:59):

Will this be the same when TAF involves an intense thunderstorm. I guess the stakes are different.

-