Background for implementing RPS in MET (22Oct 19)

RPS Definition and Purpose

From International Research Institute (IRI) verification document:
"The ranked probability score (RPS) measures the squared forecast probability error and therefore indicates to what extent the forecasts lack success in discriminating among differing observed outcomes, and/or have systematic biases of location and level of confidence. Thus, the score reflects the degree of a lack of discrimination, reliability, and/or resolution."

From verification FAQ web page (WMO):
RPS "measures the sum of squared differences in cumulative probability space for a multi-category probabilistic forecast. Penalizes forecasts more severely when their probabilities are further from the actual outcome. Negative orientation - can 'fix' by subtracting RPS from 1. For two forecast categories the RPS is the same as the Brier Score."

From "met-learning module" prepared by L. Wilson and P. Nurmi:
"The rank probability score is equivalent to the Brier score, but measures accuracy of probability forecasts when there are more than two categories."

From Wilks (2011):
"The ranked probability score is essentially an extension of the Brier score (Equation 8.36) to the many-event situation. That is, it is a squared-error score with respect to the observation 1 if the forecast event occurs and 0 if the event does not occur. However, in order for the score to be sensitive to distance, the squared errors are computed with respect to the cumulative probabilities in the forecast and observation vectors."

RPS Computation

Forecast form: Typically, categorical probabilistic forecasts are provided in the form of probabilities associated with each forecast category (in a vector form, essentially). For example, the discrete probabilities of below-normal, normal and above-normal temperatures in a seasonal forecast might be (0.2, 0.5, and 0.3) [Note that the values must sum to 1]. However, the easiest formulation for RPS is based on *cumulative probabilities*. In this example, the cumulative probabilities would be (0.2, 0.7, 1). [Note that the last category will always be assigned a value of 1 in this formulation (we assume that the categories are collectively exhaustive, so this is a necessary feature)].

Of course, this can become pretty complex if there are a lot of categories, as there seem to be for some of the CPC applications where they apparently are suggesting 19 subdivisions.

Observation form: Analogously to the forecasts, a vector of observation occurrences is needed. For the 3-category example in the case shown above, the observed event might be in the middle category. In this case, the discrete probabilities associated with each category would be (0, 1, 0) and the cumulative probabilities would be (0, 1, 1). [Note that – as with the forecasts – the last category will always be 1. Also note that the categories must be ordered, unlike some multi-category forecasts – such as "weather type", for example.]

Equations:

The simplest formulation of RPS is based on the CDFs:

$$RPS = \frac{1}{K-1}\sum_{k=1}^{K}(CDF_{fcst,k} - CDF_{obs,k})^2$$

where K is the number of categories and the CDF values are the cumulative probabilities – from the matched forecast and observation vectors.

In essence, *the RPS measures the difference between the distribution of forecasts and the distributions of observations across the categories*. This is analogous to the CRPS for evaluation of continuous probability distributions.

Example:

This example is from the "met-learning.eu" module developed by Pertti Nurmi and Laurie Wilson and focuses on 3-catgegory precipitation forecasts, based on real forecasts and observations. In this case, the three categories are <0.2 mm, 0.3-4.4 mm, and >4.4 mm. The following table shows some of the forecasts and outcomes.

Day	Observed rain	Obs category	Pr(Cat1)	Pr(Cat2)	Pr(Cat3)
1	0	1	0.7	0.3	0.0
2	0	1	0.9	0.1	0.0
3	0	1	0.9	0.1	0.0
4	0	1	0.8	0.2	0.0
5	0	1	0.8	0.2	0.0
6	0	1	0.9	0.1	0.0
7	1.1	2	0.6	0.4	0.0
8	0.9	2	0.3	0.4	0.3
9	0	1	0.3	0.4	0.3
10	0	1	NA	NA	NA
11	2.2	2	NA	NA	NA
12	0	1	0.8	0.2	0.0
13	1.2	2	0.8	0.2	0.0
14	6.0	3	0.0	0.4	0.6
15	2.3	2	0.3	0.7	0.0
…	…	…	…	…	…

Converting these forecasts to the cumulative probability form results in the following table:

Day	Observed rain	Obs category	p1	p2	p3	CDF(obs, k=1,2,3)	CDF(fcst, k=1,2,3)	RPS contribution
1	0	1	0.7	0.3	0.0	1,1,1	0.7,1,1	0.045
2	0	1	0.9	0.1	0.0	1,1,1	0.9,1,1	0.005
3	0	1	0.9	0.1	0.0	1,1,1	0.9,1,1	0.005
4	0	1	0.8	0.2	0.0	1,1,1	0.8,1,1	0.020

5	0	1	0.8	0.2	0.0	1,1,1	0.8,1,1	0.020
6	0	1	0.9	0.1	0.0	1,1,1	0.9,1,1	0.005
7	1.1	2	0.6	0.4	0.0	0,1,1	0.6,1,1	0.180
8	0.9	2	0.3	0.4	0.3	0,1,1	0.3,0.7,1	0.090
9	0	1	0.3	0.4	0.3	1,1,1	0.3,0.7,1	0.290
10	0	1	NA	NA	NA	1,1,1	NA	NA
11	2.2	2	NA	NA	NA	0,1,1	NA	NA
12	0	1	0.8	0.2	0.0	1,1,1	0.8,1,1	0.020
13	1.2	2	0.8	0.2	0.0	0,1,1	0.8,1,1	0.320
14	6.0	3	0.0	0.4	0.6	0,0,1	0,0.4,1	0.080
15	2.3	2	0.3	0.7	0.0	0,1,1	0.3,1,1	0.045
…	…	…	…	…	…	…	…	…

Accumulating RPS scores across time and space

MET computes scores for individual times across a grid (or across points); thus the RPS values MET will compute represent the overall performance of the categorical probability forecasts across that set of points at a single time.

An important issue with this approach is that combining results across the different climatologies represented by the sub-regions in the domain may lead to results that are not very meaningful (see paper by Hamill, 2006). Thus, ideally RPS will be applied to meaningful/homogeneous sub-regions. In particular, applying local climatological thresholds across the domain (e.g., the 0.3, 0.4, 0.3 categories used in the standard monthly/seasonal forecasts; alternatively, the 19 quantiles that CPC apparently uses in some of the model evaluations) can alleviate this issue.

Issues regarding computing statistics across large spatial domain – e.g., combining scores from a large set of stations in different climatological regions

- Discussions by Hamill and Juras (QJRMS 2006)
    - Climatological differences can lead to an indication of "skill" where none exists
    - Examples
        - 2 islands with opposing climatologies: Skill arises from forecasts' ability to distinguish the climatologies of the two locations, not from being able to adequately forecast weather/climate conditions at the two islands
- Solutions to this issue include
    - Evaluating across homogeneous samples
    - Moving variables to a distribution framework (e.g., looking at quantiles rather than raw variables)

Process used by EMC (from Bin Bin):

For specific forecast hour, domain, and field:

- Loop over all grid points:
- At each grid point,
    - Call RRPS( ) sub
    - Input: 20 member fcst, 10 clim bins, and 1 analysis
    - *Note*: RPS is accumulated and averaged over all clim bins in rrps ( ) sub
    - *Output*: Bin averaged RPS, for fcst and clim, respectively
    - Accumulate RPS, for fcst and clim, respectively
    - End of grid loop
- **Average RPS over all grids (i.e., they are computing it at each point across time, then averaging across the grid)**
- Divided by total number of grid), for fcst and clim, respectively
- Compute RPSS:
    - Using fcst RPS and clim RPS
- Final output to VSDB file: grid-averaged RPS for fcst, RPS for clim, and RPSS

Also from NCEP:

The data are in 32-bit floating point binary, and contain 19 records (probability of exceeding the 1st percentile, 2nd percentile, ..., 99th percentile). These are the percentiles if you need them:

[1, 2, 5, 10, 15, 20, 25, 33, 40, 50, 60, 67, 75, 80, 85, 90, 95, 98, 99]

This many categories doesn't make much sense statistically… Is there a possibility of reducing the number?

Alternative formulations

1. PDF form of RPS

The CDF form of RPS shown above is the easiest to write down, but it does require creating a CDF from the raw categorical probabilities. This step can be incorporated into the computation of RPS, as shown in the following alternative RPS formulation:

ADD PDF EQUATION HERE

2. Weigel de-biased formulation

Weigel (MWR, 2008) used a "de-biased" form of the RPS, to take into account biases that arise from using a small sample size. In their case, it was because their system only produced 5 hindcast forecasts (i.e., using 5 ensemble members), in addition to the actual forecasts with 51 members. I'm not sure how widely this de-biased form is used, but it is not too complicated to compute, if it would be useful for MET users.