Combining Probabilities

Lagging and blending percentile data

Currently when a new ensemble forecast is produced then, provided it is a higher resolution ensemble, we replace the percentile values for that forecast period with the values created from the new ensemble forecast. However what we would like to do is find a method similar to the method currently used by the deterministic Bestdata feed to lag and blend forecasts from different models together.

Currently in the deterministic Bestdata feed we have

$$x_{blended} = w_1 x_{old} + w_2 x_{new}$$

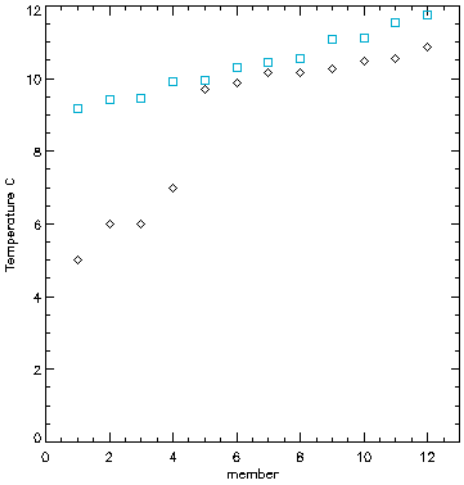
where the weights w_1 and w_2 depend on which model the forecast is from and the forecast period.

Here we are looking at the special case where we are combining two ensembles together. A separate method will be described for the case where we want to add a deterministic model to an ensemble.

To illustrate the proposed method we shall start with an example.

Suppose for a particular site and forecast period we have two ensemble forecasts (ENS1 , ENS2) with 12 members.

member	ENS1	ENS2
1	5.00	9.18
2	6.00	9.42
3	6.00	9.45
4	7.00	9.91
5	9.72	9.96
6	9.89	10.30
7	10.15	10.45
8	10.16	10.55
9	10.26	11.08
10	10.49	11.12
11	10.56	11.54
12	10.86	11.74



Calculation of the percentile values for each ensemble and the combined ensemble is an easy calculation.

Percentile	cdf1	cdf2	combined
5%	5.55	9.310225	6
10%	6	9.41908	6.3
20%	6.2	9.53888	9.322
30%	7.81714	9.92814	9.69592
40%	9.78904	10.09732	9.92356
50%	10.01695	10.374	10.1535
60%	10.1548	10.5068	10.2944
70%	10.23	10.9197	10.4931
80%	10.4416	11.1086	10.681
90%	10.5536	11.4994	11.1049
99%	10.828	11.72089	11.69677

However in practice to recalculate the combined percentiles we would need to store every ensemble that goes into the combined percentiles, extract them and calculate the combined percentile values. With two ensembles this isn't a problem but to use lagging and blending properly we would need all the models and forecast-steps, this is not practical.

We need a way to calculate the combined percentiles from the percentiles from the individual ensembles.

The obvious method of just taking the average of the percentiles is not going to work

Percentile	cdf1	cdf2	average	combined
5%	5.55	9.310225	7.430113	6
10%	6	9.41908	7.70954	6.3
20%	6.2	9.53888	7.86944	9.322
30%	7.81714	9.92814	8.87264	9.69592
40%	9.78904	10.09732	9.94318	9.92356
50%	10.01695	10.374	10.19548	10.1535
60%	10.1548	10.5068	10.3308	10.2944
70%	10.23	10.9197	10.57485	10.4931
80%	10.4416	11.1086	10.7751	10.681
90%	10.5536	11.4994	11.0265	11.1049
99%	10.828	11.72089	11.27445	11.69677

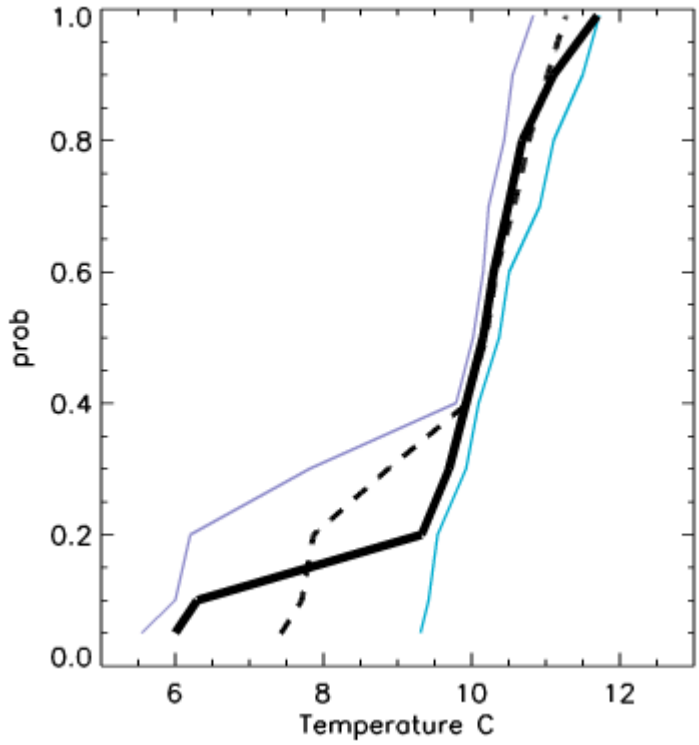


Figure 1
The blue/purple lines are the cdfs of the individual ensembles. The thick line is the combined. The dotted line is the average

This is because we are trying to combine the probability distributions, the percentiles are points fixed within the distribution dependent on the distribution, so in the combined probability distribution the outlining percentile values will reflect the change in the range from combining the two ensembles, the central percentiles will move to reflect the changes in the density of the combined pdf.

What we need to do is combine the probability spaces not the percentile values. To do this we start by calculating where the percentile values from each ensemble would exist in the others probability space.

For simplicity we use linear interpolation to calculate the revised probability i.e. for each percentile in one probability space $P(X_1 < x_{val})$ we take the percentile value above and below in the other probability space and work out what the percentile value would be in that space

$$P(X_2 \le x_{val}) = P(X_2 \le x_{above}) - \frac{x_{above} - x_{val}}{x_{above} - x_{below}}(P(X_2 \le x_{above}) - P(X_2 \le x_{below}))$$

For values below the percentile range the probability has been set to 0 and above it has been set it 100% however an appropriate curve could be fitted to more accurately reflect the extremes of the pdf.

Temperature	prob space 1	prob space 2
5.55	5%	0%
6	10%	0%
6.2	20%	0%
7.81714	30%	0%
9.310225	37.57%	5%
9.41908	38.12%	10%
9.53888	38.73%	20%
9.78904	40%	26.43%
9.92814	46.10%	30%
10.01695	50%	35.25%
10.09732	55.83%	40%
10.1548	60%	42.08%
10.23	70%	44.79%
10.374	76.80%	50%
10.4416	80%	55.09%
10.5068	85.82%	60%
10.5536	90%	61.13%
10.828	99%	67.78%
10.9197	100%	70%
11.1086	100%	80%
11.4994	100%	90%
11.72089	100%	99%

We can then combine the probabilities together using which ever weighting we would like

$$P(X_{blended} \leq x) = w_1 P(X_1 \leq x) + w_2 P(X_2 \leq x)$$

So for example if we take $w_1 = 0.5$ and $w_2 = 0.5$

Temperature	prob space 1	prob space 2	combined prob space
5.55	5%	0%	2.5%
6	10%	0%	5%
6.2	20%	0%	10%
7.81714	30%	0%	15%
9.310225	37.57%	5%	21.29%
9.41908	38.12%	10%	24.06%
9.53888	38.73%	20%	29.36%
9.78904	40%	26.43%	33.21%
9.92814	46.10%	30%	38.05%
10.01695	50%	35.25%	42.62%
10.09732	55.83%	40%	47.91%
10.1548	60%	42.08%	51.04%
10.23	70%	44.79%	57.39%
10.374	76.80%	50%	63.40%
10.4416	80%	55.09%	67.54%
10.5068	85.82%	60%	72.91%
10.5536	90%	61.13%	75.56%
10.828	99%	67.78%	83.39%
10.9197	100%	70%	85%
11.1086	100%	80%	90%

11.4994	100%	90%	95%
11.72089	100%	99%	99.5%

The final step is to recalculate the percentiles in the new combined probability space. Again we use linear interpolation to calculate the values

$$x_{val} = x_{above} - \frac{P(X \le x_{above}) - P(X \le x_{val})}{P(X \le x_{above}) - P(X \le x_{below})}(x_{above} - x_{below})$$

Perecentile	value
5%	6
10%	6.2
20%	9.004785
30%	9.580121
40%	9.965977
50%	10.13569
60%	10.2924
70%	10.47143
80%	10.70911
90%	11.1086
99%	11.69628

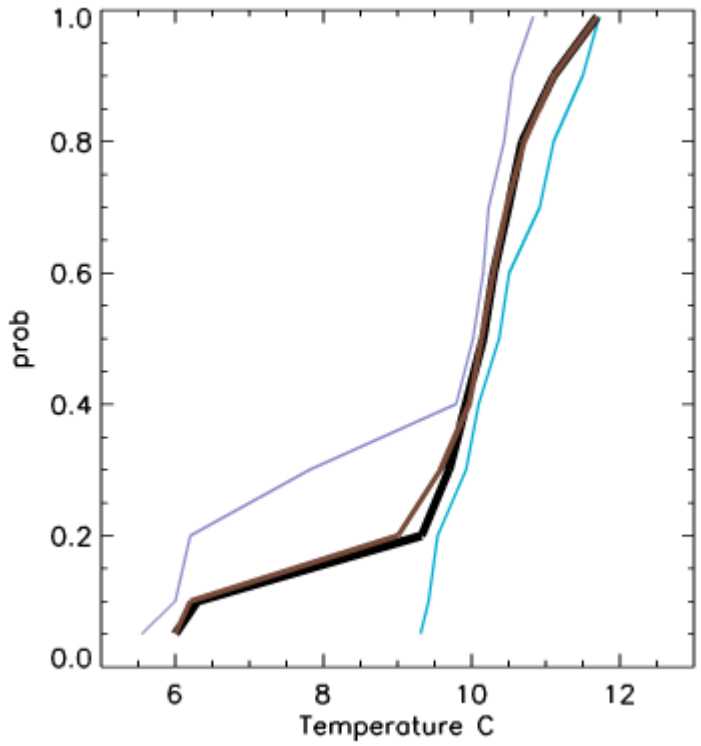


Figure 2

The blue/purple lines are the cdfs of the individual ensembles.
The thick line is the combined (calculated from the individual ensemble members).
The brown line is the cdf (calculated from the percentiles of the ensembles using the method described above)

The method does not perfectly match the values calculated using the individual ensembles and we couldn't expect it to as some of the information is lost when the ensemble is reduced to percentiles. However it does enable as to combine ensembles with any weight we would like. It is also flexible in that we can use any percentile values (note the bottom percentile was 5% the top 99%). Obviously the more percentile values used (especially at the extremes) the better.

We are currently running test to see how this method performs in practice.