In the last section, we analyzed the meteorological data ensemble, that is using different meteorological inputs and we found that the ensemble mean was better than any one of the individual calculations. There is another approach to this, instead of doing an ensemble mean of all the members, we can do an ensemble mean of the independent members, and this is called an ensemble reduction technique. What that implies is we want to eliminate members that are redundant with each other, and what I mean by redundant is if 2 simulations or 2, 2 ensemble members are very similar to each other, we want to eliminate one of them. So that all the ensemble members that we are going to compare to each other should be as independent as possible. And this gives us a better representation of the possible uncertainty or variance that we might see in the possible solutions.

The explanation for this is, for instance, if we have an uneven selection of ensemble members. Let's say we have 10 members, and five of those members consist of variations using the same meteorological data set, perhaps physics variations, and then the other five, the remaining five members represent, say different meteorology. If that one group of five ensemble members that all come from the same meteorological data set and that meteorological data set is superior to any of the others, then the ensemble weight will be toward those five, and then the overall ensemble mean is not really representative of the variance that we get over the range of members, but it would be dominated by this one member where all the results of the same. If that model were the best model, then of course the overall results would be good. But if that particular meteorological model were the worst and it had all the members, we would be biasing the results downward. So to reduce, or to determine which members are redundant, which members are very close to each other, the ensemble reduction technique relies on a metric, a statistical metric and it could be, and I think in this case it will be, the normalized mean square, for instance. And so the analysis compares every output file with every other output file, and those files that have the closest, the smallest normalized main square error, those pairs of files with the smallest normalized mean square, are then considered to be the redundant pair and one of them, you know, would be eliminated.

So this is the reduction process and it relies upon the fact that we've completed the calculations in the last step, the statistical calculations for the five data sets, which means that in your working directory you have hysplit_DATEM one through six, the 6 being the ensemble mean. So we will temporarily remove number six because we want to create a new ensemble mean that only consists of the independent members rather than all the members. You could delete it but I want to use it again for some other calculations.

Since we'll be using a wild card I'm just going to take out this text here, and I think that's all we need to do. Well one other thing, let's also do that to the data file, the original data file here. So I just should make this hysplitX, so that we know that this is experimental one that we did before.   And we should do this same here.   And now we're going to go ahead and process those.   So the way to do that is to first make sure that the wildcard is correct. Actually I’m not sure … that would be hysplit2, and then we're going to open the display, ensemble menu, again, but this time we're going to do reduction.   This will be the wild card for the selection of the files, these are the DATEM formatted files that were produced by the statistical analysis that was done in the previous section. If you want take a look at one of those files, which I should've done when I had this open, it just gives the results of the calculation, the HYSPLIT calculation in DATEM format for the CAPTEX samples.   And we need to define the measured data file, and then we're going to do the reduction, and that's just apply.   It's pretty simple. It doesn't take long for the number of data sets we had.

And we can see here that the combination with the minimum mean square, that’s what we're looking for, is right here, so comparing, when you compare all ensemble members to each other, the three that have the least minimum mean square error are members 2, 3 and 4.   So we should create now a statistical analysis of ensemble members 2, 3 and 4 as an ensemble mean.   So the first thing we need to do is go into the utility menu, in binary file merge.   So we want to create a mean, an ensemble mean concentration of members 2, 3, and 4.   Now we are ready have an INFILE, so all we need to do is edit that INFILE to include only members 2, 3, and 4.   Now we will create a new output file, and I'm actually not going to reuse the number 6, which was the ensemble mean of all of them, but I'm going to create a new one called 7, and since it consists of three members, to get the mean the multiplier would be 0.333 and we're to process those files, and we should now have number 7, that we just created. And I'm going to put back number 6, so this was the ensemble mean of all the members, we don't need to do this.

And the next step would be to re-compute the statistics. So that's display, ensemble, statistics, and I think everything stays the same, and that is we're using these measured data, the wild card name is hysplit2 and let's make the output. Well, we don't actually need to, we can overwrite some, because we added the other ensemble mean. So now you can see what we have, and so we have the original five members, we have the ensemble mean for all five members, and we have the ensemble mean for the three independent members, and for the three independent members the metric that was used was the mean square and naturally they have the lowest mean square error. But it's not necessarily true for all the other metrics, so the correlation is actually higher, well for the correlation and the mean square error are better for the three member than the five member. However the biases are slightly less and there is slightly less overlap, and the distributions are not quite as good, but the ranks are almost identical between these two.

But if you are looking to find independent members that 's the reason for using this reduction technique. The practical application of a reduction technique would be several. One example I might give you, is let's say you are doing forecasts, and you do have verification data, and you might have verification data in a sense that it might be air quality data that is collected daily and updated everyday. So you would run your meteorological ensemble for the day, for the archival day where you already have measurements, and you do an ensemble reduction to find out which ensemble members give you the best fit with the observations, and then you would use those ensemble members in the forecast for the following days, assuming that the ensemble members that gave you the best analysis yesterday, will give you the best forecast tomorrow. So this is just one approach of doing ensemble reduction. And there is more detail about this technique discussed in a publication which I would refer you to here at the end of the tutorial section on ensemble reduction techniques.

And this concludes our, section 13 on concentration uncertainty, and next we will do an exercise.