Managing Big Reanalysis Data Sets: Use of EOFs for Data Compression

Huug van den Dool
Climate Prediction Center

In principle EOFs are ideal for data compression. Yet they are not much used for this purpose. This may be in part because most EOF subroutines determine EOFs as the eigenvectors of the covariance matrix. For a typical CFSR global data set at full space/time resolution (~750,000 by 250,000) it is nearly impossible to calculate the covariance matrix because it takes too much CPU, even for a single variable.

However, unbeknownst to most people in meteorology/climatology, there are iterative methods that do not require the covariance matrix. They are much cheaper and also much simpler to execute, and as long as we need only a limited number of EOFs (a few hundred at most) this approach works. By example we will show a) that the iterative method yields the same results (on low resolution data) as the standard covariance matrix approach, and b) that we succeeded in calculating the first 200 EOFs of a global single variable daily CFSR data set at full spatial resolution. People can thus store some of the CFSR data sets at a fraction of the memory on their devices, and still reconstruct the data at full resolution for any calculations. One does have to truncate, and give up 1 or 0.1% explained variance (this is a choice). For detailed data sets like rainfall the iterative method will fail because too many EOFs are required. But most large scale fields can be compressed satisfactorily.

Corresponding Author:

Name: Huug van den Dool
Organization: Climate Prediction Center
Address: 5200 Auth Rd
Camp Springs, MD 20746
USA