Introduction

Air mass classification has become an important area in synoptic climatology, simplifying the complexity of the atmosphere by dividing the atmosphere into discrete similar thermodynamic patterns. However, the constant growth of atmospheric databases in both size and complexity implies the need to develop specific classifications to extract the essential information adapted to climate study. In particular, the classification of atmospheric columns homogeneous in temperature and humidity is of great importance in inverse problem where climate variables as well as the 3-dimensional structure of the atmosphere can be estimated from satellite data via inverse radiative transfer models. Such models need to be initialized by a priori information as thermodynamic vertical profiles and surface variables which usually come from radiosonde or raumyns. Here, we present a robust air mass unsupervised and supervised classification methodology of large space-time atmospheric datasets, providing not only the resulting classes but also the a posteriori probability to belong to each class. This methodology follows a probabilistic point of view: both classification process and data are probabilistic. It is applied to thermodynamic profiles (temperature and dew point temperature) coming from ECMWF reanalyses. These data are gridded in latitude, longitude and vertical layers, homogeneously distributed over the Earth and over a decade (2000-2009).

1. Methodology

(a) Average temperature in the 800-425 hPa layer (in K)

(b) Total column water vapour (in precipitable cm)

From classical data (T, H2O) profiles to probabilistic data (CDF)

Temperature profile

Clustering a Global Field of Atmospheric Profiles by Mixture Decomposition of

Clustering: An unsupervised classification methodology of large space-time atmospheric datasets

2. Unsupervised classification of atmospheric situations into air masses homogeneous in temperature and humidity

- Only 3 covariance matrix models lead to relevant air mass spatial regions with:
  - inverse hyperbolic model (i.e., ITCZ, tropical center continental polar air mass, depressions);
  - neither low nor high equability regarding the size of the clusters (ii).
- Here, use of the hyperbolic models Σ = κI and Σ = κI + J which both assume isotropic dispersion (equal or variable spread on the surface, respectively) of each CDF values samples related to a given cluster k in the D-dimensional space.
- [EM] Σ = κI leads to a partition which is similar to the one obtained with the widely used k-means algorithm (Fig. 4).
- [EM] Σ = κI + J detects more accurately tropical air masses compared to [EM] Σ = κI (Fig. 5, 6, 7), due to a higher relative influence of humidity over temperature on the classification regardless the amount of humidity expected patterns (i).
- The classification method is quite consistent with the number of clusters (iii), meaning adding successively one cluster does not drastically change most of the clusters, in particular between 7 and 8 clusters (Fig. 4, 5) or between 12 and 13 clusters (Fig. 5).

Classification of the training dataset (15th day of Jan/Apr/Jul/Oct 2005-2009 at 00h and 12h UTC) with no spatial sampling

Classification of new atmospheric situations

3. Sensitivity in time and space of the classification

Little sensitivity of the classification to the choice of the spatial-temporal sampling of the training dataset (iv):
- as soon as 4 months representative of each season are considered;
- for a spatial sampling step 1 not exceeding 7 (i.e., 1 grid point out of 7 consecutive ones = 7/3° × 7/3° ≈ 12° in both longitude and latitude).
- Classification sensitivity to spatial sampling step 5 and starting grid point (among the 5 possible ones) through misclassification rate (%)

4. Supervised classification of new atmospheric situations

- Temporal and spatial sampling of the training dataset used for the classification of new atmospheric situations over several years:
  - 15th day of Jan/Apr/Jul/Oct 2005-2009 at 00h and 12h UTC;
  - 12° × 12° spatial sampling with random first grid point.
- Few differences between supervised and unsupervised classifications, located at the transition between the classes.

5. Importance of a posteriori probabilities

- Contrary to k-means, EM provides the a posteriori probabilities of belonging to each class (Fig. 4, 5), located at the transition between the classes (Fig. 6: air masses are rather well separated).

Conclusion

The classification method presented here complies with the five properties (i) to (v) introduced by Hath (1996), Hath et al. (2008) to assess the quality of a classification. EM generalizes the widely used k-means algorithm by providing, in addition, the a posteriori probabilities of belonging to each class for each atmospheric situation and thus the corresponding error probabilities as well, which can be used for many applications (e.g., adding transition classes, improving a priori information in remote sensing).

Based on temperature and dew point temperature variables only, the method is applicable to most atmospheric datasets used by the atmosphere science community, such as raumyns, radiosonde measurements or satellite data. Depending on the intended objective, other variables could be added, especially dynamic variables to help monitor air mass movement.

The method could be easily adapted to evaluate global circulation models and study climate variability and potential changes at different spatial and temporal scales.