7.2 REGIONAL-SCALE WIND FIELD CLASSIFICATION EMPLOYING CLUSTER ANALYSIS

Lee G. Glascoe (glascoe1@llnl.gov), Ronald E. Glaser, Hung-Neng S. Chin, Gwendolen A. Loosmore
Lawrence Livermore National Laboratory
Livermore, CA 94550 USA

1. INTRODUCTION

The classification of time-varying multivariate regional-scale wind fields at a specific location can assist event planning as well as consequence and risk analysis. Further, wind field classification involves data transformation and inference techniques that effectively characterize stochastic wind field variation. Such a classification scheme is potentially useful for addressing overall atmospheric transport uncertainty and meteorological parameter sensitivity issues.

Different methods to classify wind fields over a location include the principal component analysis of wind data (e.g., Hardy and Walton, 1978) and the use of cluster analysis for wind data (e.g., Green et al., 1992; Kaufmann and Weber, 1996). The goal of this study is to use a clustering method to classify the winds of a gridded data set, i.e, from meteorological simulations generated by a forecast model.

2. PREDICTED WIND FIELDS

The predicted regional scale wind fields were generated using the Coupled Ocean Atmosphere Prediction System (COAMPS) model (Hodur, 1997) on three-hourly intervals for several altitude levels. We employed the ADAPT atmospheric data assimilation model (Sugiyama and Chan, 1998) on the COAMPS output to provide mass-consistent three-dimensional time-varying wind fields for the NARAC Langrangian particle tracking code, LODI (Nasstrom et al., 2000). The LODI code was used for all dispersion calculations.

Wind fields for the particular regional location of concern vary on both an hourly basis and on a seasonal basis (see Figure 1 and Figure 2). The winds tend to be faster and more westerly in winter, due to the dominance of synoptic forcing, and tend to be slower and more variable in the summer.

Dispersion simulations using these wind fields were conducted for event planning and consequence assessment purposes. The ensemble of dispersion runs are shown below for an instantaneous January release at 3pm (Figure 3) and for an identical July release at 3pm (Figure 4). While obvious patterns can be discerned, a quantitative method for classifying the wind magnitude, direction and duration is desirable.

3. WIND CLASSIFICATION METHODOLOGY

In cluster analysis the objective is to divide a set of observations (here the collection of gridded wind field data at various times of the year) into groups or clusters in such a way that most pairs of observations which are placed in the same cluster are more similar to each other than are pairs of observations which are placed in two different clusters. Because components are measured in the same units (m/s) it is reasonable to use Euclidean distance as a measure of similarity. The distance between the pair of observations corresponding to times t and τ is therefore

$$d_{t\tau} = \left[\sum_{j=1}^{N} \left[\left(u_{jt} - u_{j\tau} \right)^2 + \left(v_{jt} - v_{j\tau} \right)^2 \right] \right]^{1/2} \qquad (1)$$

where N denotes the number of spatial positions in the grid. In equation (1) we are consider the 2-dimensional hindcast wind velocity vector (u, v) parallel to the surface of the earth. The restriction to two velocity dimensions is motivated by computational convenience and is unnecessary.

We use a K-means clustering method which fixes the number of clusters, K, and divides the observations into K clusters in such a way that the total sum of squared Euclidean distances between observations and their respective cluster centroids is minimized. The minimization for our dataset was performed with Matlab software. There is no consensus in the statistical community on a method to select an appropriate value for the number of clusters, K. We have used the silhouette measure of Kaufman and Rousseeuw (1990) which is implemented in Matlab.

The relative frequency of occurrence of a given cluster can be estimated by the proportion of data points (times of the year) assigned to this cluster. Moreover, the wind field variation within the cluster can be characterized by parameter estimation over the associated times. For example,

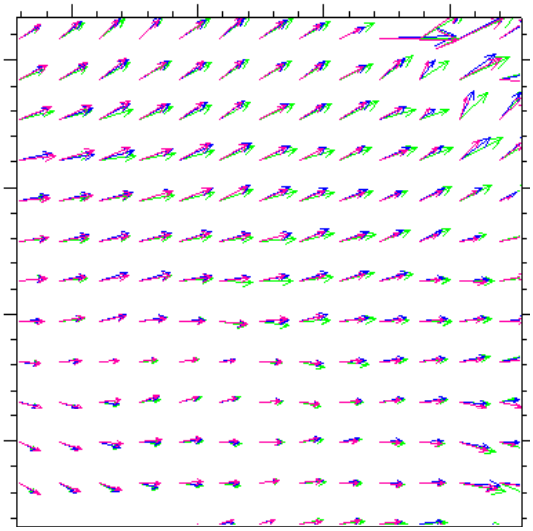

Figure 1. Time-varying wind-field on a January afternoon at 406m AGL (3pm, 4pm, 5pm). Domain shown is ~200km on a side; center magenta wind vector is ~ 5 m/sec.

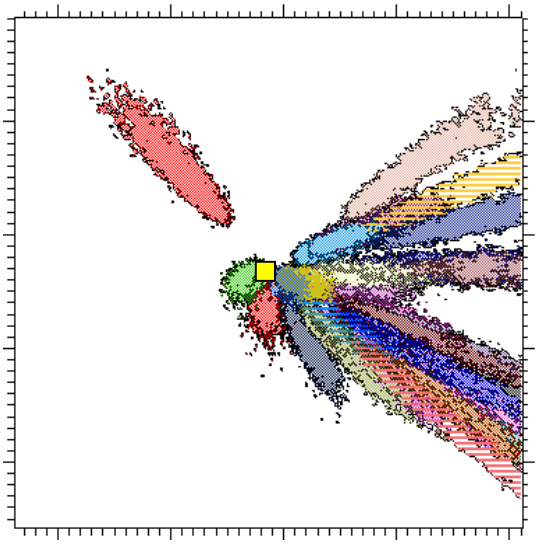

Figure 3. Ensemble of dispersion results for 30 days in January (45km x 45km domain shown, yellow box marks release point). Illustrated is a 30 day ensemble of the deposition of 100-200 micron particles instantaneously released aloft from the center location at 3pm.

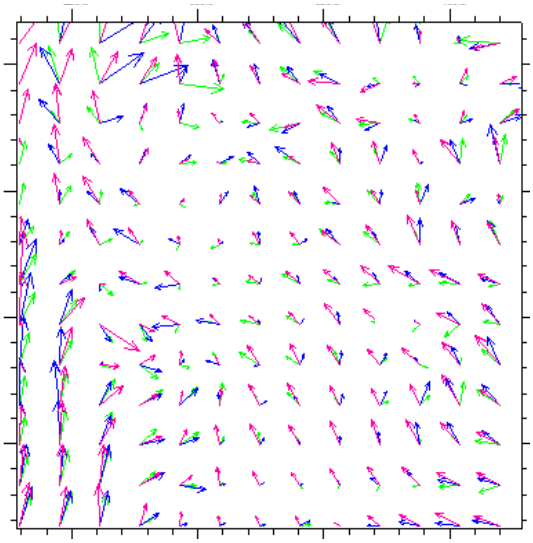

Figure 2. Time-varying wind-field on a July afternoon at 406m AGL (3pm, 4pm, 5pm). Domain shown is ~200km on a side; center magenta wind vector represents ~2 m/sec.

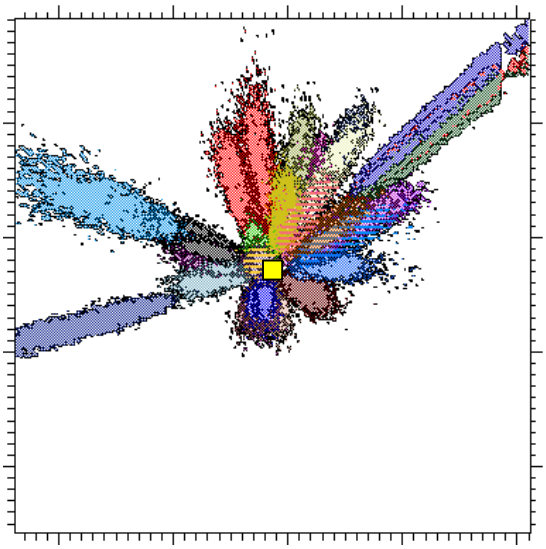

Figure 4. Ensemble of dispersion results for 30 days in July (45km x 45km domain shown, yellow box marks release point). Illustrated is a 30 day ensemble of the deposition of 100-200 micron particles instantaneously released aloft from the center location at 3pm.

the 2-dimensional velocity vector $(U_j, V_j)^{(c)}$ at a particular location *j* can be modeled for cluster number *c* ($1 \le c \le K$) by a bivariate distribution with mean, standard deviation, and correlation parameters $(\mu_u, \mu_v, \sigma_u, \sigma_v, \rho_{uv})_j^{(c)}$, which are estimated by their sample counterparts; for example,

$$\hat{\mu}_u^{(c)} = \frac{1}{m_c} \sum_{t \in T_c} u_{jt} \qquad (2)$$

where T_c is the set of points (times) assigned to cluster *c*, and m_c is the size of this set, i.e., the number of points in the cluster. From this we may obtain, by assuming bivariate normality, estimated probability contours at levels of interest, say 50%, 75%, and 90%. The level γ probability contour is an ellipse within which an expected percentage, γ, of values $(U_j, V_j)^{(c)}$ will fall. The location, orientation, and shape of the ellipse depend on the five parameter values.

The clustering method used is graphically described in Figure 5 and Figure 6 for a simple example with 9 spatial readings at 6 times (t1 through t6). As an example, if three clusters are chosen, i.e., a certain maximum distance *d* is allowed for dissimilarity, then the six time readings reduce to 3 clusters each having a corresponding average wind components and measures of variance and correlation, $(\mu_u, \mu_v, \sigma_u, \sigma_v, \rho_{uv})_j^{(c)}$.

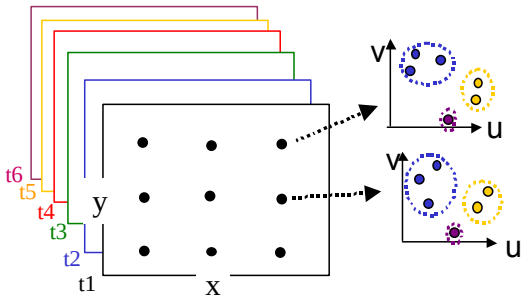

Figure 5. Example of clustering: 6 time sets of 9 spatial observations.

4. CLASSIFICATION OF 2003 WINDS

For this study we enjoyed access to a large set of hindcast wind field data and an ensemble of dispersion simulations over a specific region for every 3 hours over the entire year of 2003. Using *K*-means we partition this entire year's worth of noon and midnight data into distinct clusters which are then characterized by probability distributions that describe spatially varying wind speed and direction. For the presented analysis we chose to cluster at a single altitude of 187m to capture a release aloft, sacrificing any possibility to capture vertical wind shear in the wind classification. This was deemed acceptable as the location of concern experiences little wind shear below the planetary boundary layer depth (~400m at midnight, ~1200m at noon). The spatial observations of wind were made for every 36km x 36km over a 220km x 220km domain, i.e., for 49 spatial observations at every time reading.

The u-v data (one point for each reading) for 2003 are plotted for the center "release" location of the domain in Figure 7. Note that there are another 48 of these u-v plots diagrams to account for all of the 49 spatial locations involved. The 98-dimensional data were clustered using the *K*-means method described above with best clustering according to silhouette criteria occurring for 5 wind classes. The projected clustering for the center location is illustrated in Figure 7: a low-wind cluster (red-center), and winds from four directions (orange-SE; dark blue-SW; light blue-NW; green-NE). As discussed above, each of the 5 classes also has associated with it the characterizing parameters (μ_u, μ_v, σ_u, σ_v, ρ_{uv}) for any given location. Normal distribution probability contours can in turn be generated for each wind class (50% contours, see Figure 7; 75% contours, see Figure 8) that may be employed in sensitivity and uncertainty analysis.

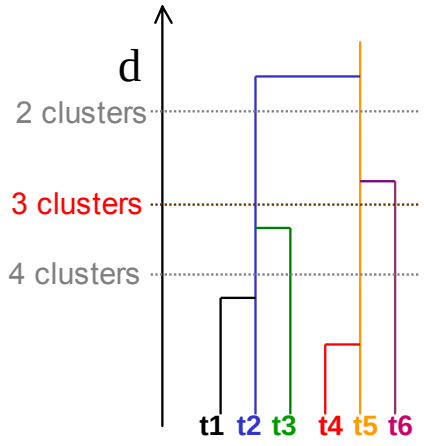

Figure 6. Example of clustering: 3 clusters group as t1, t2 & t3; t4 & t5; and t6.

The wind cluster frequency can be determined on a diurnal basis and on a longer term basis. While clustering was done for the entire year, it is instructive to demonstrate how the clustering scheme separates the frequency of wind classes in distal months January and July (Figure 9). January is dominated by westerly winds occurring about 70% of the time (northwesterly winds are particularly dominant) and with low winds occurring about 25% of the time. Westerly winds are expected as the synoptic patterns are primarily northwesterly in nature and are very dominant in the winter for this specific location. July is dominated by low winds occurring nearly 50% of the time and by southerly winds occurring nearly 40% of the time. This indicates a lack of synoptic forcing during the summer month and the dominance of local wind patterns in the region.

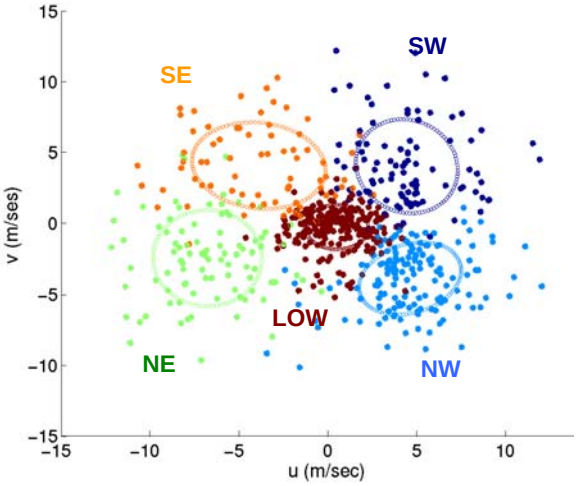

Figure 7. The 2003 u-v wind data for 12pm and 12am wind forecasts at 187m altitude for the center "release" location. Note 50th percentile contours (five clusters are best).

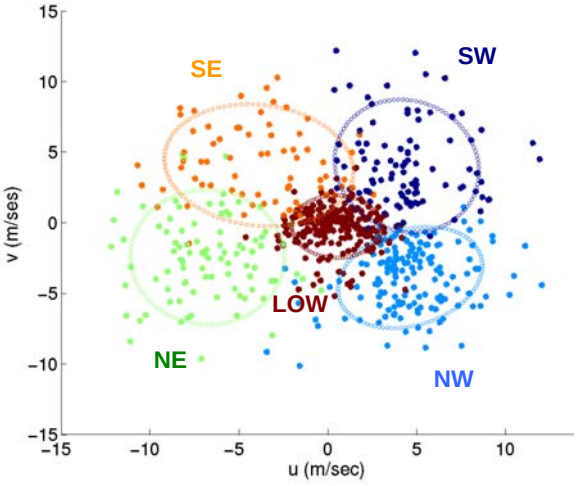

Figure 8. Same data (u and v) as in Figure 7 but with 75th percentile contours.

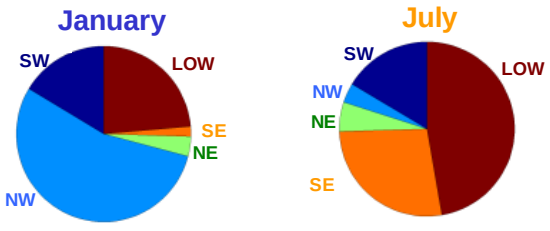

Figure 9. Wind cluster frequency during January and July.

Comparison of the frequency of occurrence (Figure 9) and the ensemble of dispersion simulations for 30 days in January (Figure 3) and for 30 days in July (Figure 4) illustrate the usefulness of the five wind clusters. The dominance of westerlies is evident in both dispersion runs and the clustering for January; the dominance of low winds and southerlies is evident in both dispersion runs and the clustering for July.

5. CONCLUSIONS

The *K*-means clustering method is effective in using a heterogeneous high-dimensional multivariate data set to create a manageable set of relatively homogeneous classes which can be characterized stochastically and employed in event planning and consequence assessments as well as in sensitivity/uncertainty analyses. The single altitude clustering of the 2003 noon and midnight gridded wind fields results in five identifiable wind classes that generally agree with an ensemble of dispersion simulations. While this example does not account for multiple altitudes and does not account for a time-series relationship for the change in wind classes, the study demonstrates the method's utility for classifying events of short duration in environments with little vertical wind-shear. The method can be modified to account for additional altitudes, a vertical velocity component, and a time history of wind class change.

6. ACKNOWLEDGEMENTS

Funding for this effort was provided by the Lawrence Livermore National Laboratory Engineering Directorate to fulfill Engineering Technology Base needs. This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48. Document UCRL-CONF-204808.

7. REFERENCES

Green, M.C., L.O. Myrup, and R.G. Flocchini, 1992: A method for classification of wind field patterns and its application to southern California. *Int. J. Climatol.*, **12**, 111-135.

Hardy, D.M. and J.J. Walton, 1978: Principal components analysis of vector wind measurements. *J. Appl. Meteor.*, **34**, 49-67.

Hodur, Richard M., 1997: The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). *Monthly Weather Review*, **125**, 1414–1430.

Kaufman L. and P.J. Rousseeuw, 1990: *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley.

Kaufmann, P. and R.O. Weber, 1996: Classification of mesoscale wind fields in the MISTRAL field experiment. *J. Appl. Meteor.,* **35**, 1963-1979.

Nasstrom, J.S., G. Sugiyama, J.M. Leone, Jr., and D.L. Ermak, 2000: A real-time atmospheric dispersion modeling system, *Eleventh Joint Conference on the Applications of Air Pollution Meteorology*, Long Beach, CA, Jan. 9-14.

Sugiyama, G. and S.T. Chan. *A New Meteorological Data Assimilation Model for Real-Time Emergency Response*. in *10th Joint Conference on the Applications of Air Pollution Meteorology*. 1998. Phoenix, AZ: American Meteorological Society.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences*, Academic Press, New York.