한국환경과학회지 제15권(제12호), 1125~1139, 2006
Journal of the Environmental Sciences

Classification of Daily Precipitation Patterns in South Korea using Mutivariate Statistical Methods

Janos Mika*, Baek-Jo Kim and Jong-Kil Park**
Meteorological Research Institute, Korea Meteorological Administration, Seoul 156-720, Korea
Hungarian Meteorological Service, H-1675 Budapest, P. O. Box 39, Hungary
**Department of Atmospheric Environment Information Engineering, Inje University, Gimbae 621-749, Korea*
(Manuscript received 22 June, 2006; accepted 2 November, 2006)

The cluster analysis of diurnal precipitation patterns is performed by using daily precipitation of 59 stations in South Korea from 1973 to 1996 in four seasons of each year. Four seasons are shifted forward by 15 days compared to the general ones. Number of clusters are 15 in winter, 16 in spring and autumn, and 26 in summer, respectively. One of the classes is the totally dry day in each season, indicating that precipitation is never observed at any station. This is treated separately in this study. Distribution of the days among the clusters is rather uneven with rather low area-mean precipitation occurring most frequently. These 4 (seasons) × 2 (wet and dry days) classes represent more than the half (59 %) of all days of the year. On the other hand, even the smallest seasonal clusters show at least 5~9 members in the 24 years (1973-1996) period of classification. The cluster analysis is directly performed for the major 5~8 non-correlated coefficients of the diurnal precipitation patterns obtained by factor analysis in order to consider the spatial correlation. More specifically, hierarchical clustering based on Euclidean distance and Ward's method of agglomeration is applied. The relative variance explained by the clustering is as high as average (63%) with better capability in spring (66%) and winter (69 %), but lower than average in autumn (60%) and summer (59%). Through applying weighted relative variances, i.e. dividing the squared deviations by the cluster averages, we obtain even better values, i.e. 78 % in average, compared to the same index without clustering. This means that the highest variance remains in the clusters with more precipitation. Besides all statistics necessary for the validation of the final classification, 4 cluster centers are mapped for each season to illustrate the range of typical extremities, paired according to their area mean precipitation or negative pattern correlation. Possible alternatives of the performed classification and reasons for their rejection are also discussed with inclusion of a wide spectrum of recommended applications.

Key Words : Cluster Analysis, Daily Precipitation Patterns, Factor Analysis

1. INTRODUCTION

As the development of observation, transmission and archiving technologies, amount of available digital information increases rapidly. For the dynamical methods and numerical weather models, governed by well-settled equations of physics, this "exponential" (i.e., self-enhancing, always faster than before, although no exact time-dependence is known by the authors) development can be used in a fairly straightforward way. Of course, this development requires much innovation in data assimilation, numerical model development, subgrid-scale parameterization, etc.

As every statistical operation requires a minimum size of homogeneous samples, moreover, the more complex the method, the larger the required minimum, this contradiction is a limiting factor to achieve as fast accumulation of new statistical results, as fast the available data. More specifically, the problem is that the exponential increase of information in space and also of potentially available data on relevant environmental indicators express strong pressure on the size of homogeneous samples that are able to increase just in the linear pace, as time passes and the data are collected.

Corresponding Author : Baek-Jo Kim, Meteorological Research Institute, Korea Meteorological Administration, Seoul 156-720, Korea
Phone: +82-2-836-0687
E-mail: bjkim@metri.re.kr

One can help to resolve this contradiction by searching for most informative components of the relevant information expressing them in the less possible abstract dimensions of independent coefficients or qualitatively different classes of events. This operation, however, may cause some loss of information and it requires additional professional expertise governed by current paradigms of atmospheric or related environmental sciences to decide where to terminate this signal/noise delimitation.

The present paper recommends a classification of daily precipitation patterns represented by 59 stations of South Korea. The final 14-24 classes transforming the 59 continuous variables into one discrete number are defined by cluster analysis after a preliminary factor analysis. Precipitation has been selected to demonstrate the following methodology as this variable is mainly related to meso-scale atmospheric phenomena and influenced by physical processes of even smaller scales, including microphysics of cloud droplets and crystals. Hence, deterministic computation of this atmospheric variable is rather limited compared to the requirements of medium-range weather forecasts of especially, climate change scenarios. On the other hand, this important atmospheric variable inter-relates with many environmental factors and has large impact on several sectors of economy. For these two reasons, precipitation is used in statistical research and applications quite frequently.

Of course, neither the factor nor the cluster analysis is new tool even in precipitation climatology of South Korea. Several studies have already been performed[1~5]. Our study differs from those in its scope, namely the above studies are aimed at objective classification of the stations (sub-regions), whereas this study is focused on determinating of different types of daily precipitation pattern. In other words, this study classifies the days of the available archive.

The paper is arranged as follows: Section 2 describes the data set used for the classification. Section 3 briefly specifies the selected alternatives of cluster analysis performed on the major non-correlated coefficients of the daily precipitation patterns obtained by factor analysis. Section 4 comprehends the results, sequenced for the obtained factor loadings, possible termination alternatives of the performed hierarchical clustering and detailed specification of the final classification, together with its quantitative validation. The results are discussed in Section 5 with recommended applications of the performed classification.

2. DATA

Daily precipitation data observed at 59 stations of South Korea are used, including Jeju Island. The Ullung Island as a single station is omitted in considering the results of factor analysis.

The precipitation classification is performed for 24 years between 1973 and 1996. This period (8,760 days, as the 6 leap-days were excluded) is separated into shorter sub-samples to reduce the inhomogeneity caused by the annual cycle of precipitation (Fig. 1). More specifically, area averages of the 59 stations are considered including all days, i.e. 24 values on each calendar day and also on the wet days, i.e. on days when the observed precipitation is at least 0.1 mm at the given station. Besides the averages, we also calcu-

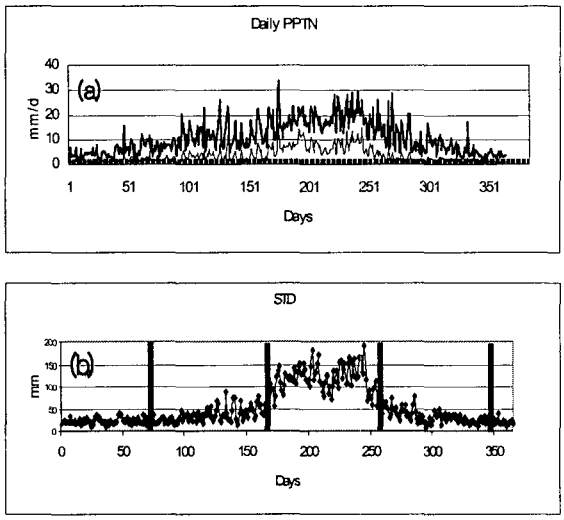

Fig. 1. Annual cycle of precipitation in 1973-1996: (a) averages for all days (thin solid line) and wet (>0.1 mm) days(thick solid line), (b) standard deviation on wet days of a station (59 stations, Ullung Island is already excluded). The defined seasons start on Julian day 349 (December 16) in winter, Julian day 75 (March 16) in spring, Julian day 167 (June 16) in summer, and Julian day 258 (September 16) in autumn.

lated the area mean value of point-wise standard deviation on the wet days.

Although three curves exhibit strong inter-diurnal variability, there is a well distinguished period in the middle of the year when all the curves are in maximum. This is the well-known summer period, connected with bi-directional march of the monsoon ("Changma") front over the Korean Peninsula and with the smaller scale convective activity including the episodic typhoons. This 3-monthly period is selected to be one season, whereas its winter opposite, exhibiting the minimum average and standard deviation, can also be clearly delimited. So, we selected four 3-monthly periods, but one should note that, according to Fig 1., each season starts by 15 days later than the general ones. In the followings, the seasons, called winter, spring, summer and autumn will always mean the sequence of the year between December 16 ~March 15, March 16~June 15, June 16~ September 15 and September 16~December 15.

Figure 1(b) shows temporal variation of the standard deviation at the 59 stations. The square of this field is the reference (100 %) being used for evaluation of the classification. Next step of classification is to separate those days of the 24 years data set, when there was no measurable precipitation at any of the 59 stations. These days are set into a separate class, a priori, since many related atmospheric and environmental fields may behave specifically in such days. This selection should be performed separately since the methodology itself, does not make this separation because the universally 0.0 mm precipitation pattern is just little different from the days in largest cluster exhibiting low amount of precipitation.

The rest of the analyses is mainly related to this 72 % of the days with measurable precipitation in one station at least. Since the proportion of stations and days with <0.1 mm is 71 %, majority of the stations and days (59%) is characterized by 0.0 mm even in this part of the sample.

3. METHODS

Classification of diurnal precipitation patterns is performed by cluster analysis applied not for the point-wise observations, but for the series of rotated loadings, computed by factor analysis. The aim of this step is not dimensional reduction, but consideration of spatial correlation of precipitation among the individual stations to make difference. Without that, the cluster analysis would take a difference in two neighboring stations with the same weight into consideration, as case of in two far-situated ones.

3.1. Factor analysis

Factor analysis is generally used for reduction of data sets, represented in a large number of stations or grid points, still keeping the essence of their common variability, by resolving the initial variables into much fewer common factors. The monograph by von Storch and Zwiers [7] gives a comprehensive theoretical overview. Factor analysis can also be used for immediate pattern classification[8], but this requires larger number of spatially distributed data (stations or grid-points) than cases (days) to be classified.

Each original variable, P_i, i=1, 2, ..., n, can be expressed as $P_i = a_{i1}F_1 + a_{i2}F_2 + a_{i3}F_3 + \ldots + a_{im}F_m$ ($m < n$), where F_j, j=1, 2, ..., m, common for each i, are the factors and a_{ij} the loadings (scores). First, it is necessary to specify whether the factor analysis is performed on a correlation matrix or a covariance matrix. We selected the covariance matrix to avoid the non-natural transformation related to standard deviation of precipitation at each station. In our view, this is also an immanent feature of the spatial pattern which is meaningful to keep before the analysis. This selection also determined that we perform a principal components analysis, as a specific realization of factor analysis.

An important question is the number of the factors (m) to retain. On this matter, many criteria have been proposed. It is noted that Jolliffe[9] states "...different objectives for an analysis may lead to different rules being appropriate". In this study, the Rule 1 or Guttman criterion is used, which determines to keep the factors with eigen-values be more than 1 and neglect the ones that do not account for at least the variance of one standardized variable.

Another vital stage is whether, or not, we should rotate the axes (factors). This process achieves discrimination among the loadings, making the rotated axes easier to interpret. In this analysis the Orthogonal Varimax Rotation is applied, which keeps the factors non-correlated. After rotation, the original precipitation P_i at station, i, is

$P_i = a'_{i1}Fr_1 + a'_{i2}Fr_2 + a'_{i3}Fr_3 + \ldots\ldots, a'_{im}Fr_m (\ m < n\ )$, where the a'_{i1}, a'_{i2}, ... , a'_{im} rotated loadings are later used for classification. The loadings are specified by the regression method to ensure the best fit of the initial data at each station.

An important feature of the non-rotated a_{ij} loadings is that they represent a decreasing order of variance as j increases. This variance is equal to the eigen-values of the analyzed covariance matrix. Moreover, the factors is the most effective set of orthogonal functions in the sense, that they "explain" the highest portion of variance retaining any fixed number of *m*. After rotation we loose this optimum feature and the variance is distributed more evenly among the retained loadings. Nevertheless, increasing sequence of loadings mean decreasing importance in explaining the variance, which is a key feature, used in cluster analysis with special emphasis. Application of rotation is not compulsory, but in our case it is also explained by the strongly skewed distribution of the loadings of first (non-rotated) factors.

Factor analysis was already used for annual precipitation of Korea by Moon [3] to classify the territory of the country. Similar interpretation is briefly exposed for the unified annual sample. The point of this application is the mapping of the maximum a'_{ij} loading at the station, i, which selects that Fr_j rotated factor for which the loading is >0.7. (According to the general experience, this threshold designates maximum one region to belong to.)

Another interpretation of the factor analysis is related to the communalities, i.e. that part of the initial variance which can be "explained" by linear combination of the retained factors and the loadings. If this mean square difference between the original and the estimated values is low for a given station, than this station exhibits large individual variations not correlated to those at the other stations.

3.2. Cluster analysis

Cluster analysis produces hierarchical clusters of cases based on distance measures of dissimilarity or similarity[10]. This method is employed to classify the diurnal precipitation patterns, represented by series of rotated factor loadings. The latter ones exhibit larger variance at the lower serial number of factors, proportionally to the explained variance, i.e. stronger difference generally occurs in most important components. Besides the 5~8 factors, explaining 81~89 % of the total variance, the analysis is also performed for that number of loadings, which commonly explain 95 % of the initial variance.

Since we do not know how many clusters to define, hierarchical joining is applied. This requires preliminary definition of distance measures for any pair of cases and specification of the algorithm, which unequivocally selects the two clusters (or single cases) to be unified at a given stage of amalgamation. The latter is based on a distance index to be minimized, which characterizes common dissimilarity.

We use Euclidean distance measure, i.e. the square root of the sum of the squared differences between the components of each case. Unification of the clusters follows the Ward's method[11], which minimizes the sum of all within-cluster variances.

The most difficult step of cluster analysis is the decision about the number of clusters to retain and interpret as the final solution. This decision should consider the retained number of classes and the efficiency of the classification. Both conditions can be considered by analyzing the agglomeration schedule, which shows the order and distances at which cases and clusters are combined into a new cluster.

If the function of distance on the number of cases exhibit sudden changes (breaks), then the clustering can be naturally terminated before such an increasing jump. One should note however, that this distance is related to the smoothed factorial representation of precipitation field, not to the total variance of the original patterns. Hence, a more established solution should be based on computation of the quality of representation for the original fields. The explained variance of the clustering, EV(k), depending on the number of clusters, k, is a key characteristic for this, defined as

$$EV(k) = \sum_{i=1}^{k} \frac{\sum_{j=1}^{N_i} \frac{1}{58} \sum_{s=1}^{59} (P_{ijs} - < P_{is} >)^2}{N_i - 1}$$

where $< P_{is} >$ is the cluster-mean precipitation at station, s, derived from all P_{ijs} values of the N_i days, that belong to the i-th cluster. For better interpretation, this explained variance is compared to the $k = 1$ version, EV(1), and the $RV(k) = EV(k)/EV(1)$ relative variance is expressed in %. The lower this ratio,

the more effective the clustering.

Since in our case both quality indices behaved rather smoothly, not allowing an optimum selection by this criterion, another point of view, i.e. the minimum number of cases in the smallest cluster, was also considered. So the selection of the final clustering is performed in three steps:

1) Candidates for termination were selected according to the N_{min} size of the smallest cluster, i.e. $N_{min} > 1$, $N_{min} \geq 5$, $N_{min} \geq 1\%$ and the last stage before $N_{min} \geq 5\%$.

2) The RV(k) relative variances were determined for the candidates. The second one was selected, as it was just slightly worse than the first one and they represented the a priori expectations about the seasonal differences: i.e. the most clusters occurred in summer and no big differences took place among the three other seasons.

3) This selection was finally corrected by decreasing the number of clusters by one in spring and summer, which gave further slight improvement in the explained variance.

The final clustering was also characterized by a modified formula, taking the differences of the cluster-mean precipitation into consideration by inversely weighting the squared deviations, as

$$WV(k) = \sum_{i=1}^{k} \frac{\sum_{j=1}^{N_i} \frac{1}{58} \sum_{s=1}^{59} (P_{ijs} - < P_{is} >)^2 / < P_{is} >}{N_i - 1}$$

This weighted variance is also standardized by the WV(1), no-clustering reference value. The idea of this alternative index is to consider if the variance is dominant in the high or low-precipitation clusters. Of course this parameter can not be applied for the dry (no precipitation) cluster.

4. RESULTS

4.1. Factor scores for further analysis

The loadings resulted by factor analysis, are just input variables for the pattern classification. The factors and the seasonal loadings are analyzed in a separate study (Mika et al, 2001). Here we focus on four aspects of these computations, evaluating:

- How detailed is the representation of the original patterns by the applied 5 ~ 8 loadings?
- Can factor analysis or rotation change the strongly skewed distribution of precipitation?
- How can we interpret the geometry of the rotated loadings, applied for regionalization?
- How can factor analysis be used to decide about inclusion of Jeju and Ullung Islands?

The retained factors and the explained variances are shown in Table 1 to illustrate the answer to the first question. The retained 5 in winter, 6 in spring and autumn or 8 in summer factors explain the 81 ~ 89 % of variance. To explain 95 % of that, we would need much more, 14 ~ 30 factors to retain. These seasonal differences correspond to the conditional means, also indicated in the Table. Considering the wet days only, the area mean precipitation is almost six times larger in summer (8.56 mm/day) than in winter (1. 52 mm/day).

Table 1. Retained original and rotated % eigenvalues of seasonal factor analysis

Factor \ Season	Winter		Spring		Summer		Autumn	
Sample Size	1,508 days		1,402 days		1,952 days		1,406 days	
Area Mean	1.52mm		4.74mm		8.56mm		2.51mm	
Explained Variance	Original %	Rotated %	Original %	Rotated %	Original %	Rotated %	Original %	Rotated %
First Factor	64	41	67	29	39	15	59	25
Second Factor	12	22	9	22	17	15	10	20
Third Factor	5	12	5	19	7	14	7	16
Fourth Factor	4	7	4	9	6	14	5	12
Fifth Factor	2	5	2	5	5	11	3	11
Sixth Factor	·	·	2	4	3	4	2	2
Seventh Factor	·	·	·	·	2	6	·	·
Eighth Factor	·	·	·	·	2	2	·	·
$\sum$ Indicated	87.5		89.0		81.0		86.4	
95% Expl.	14		15		30		17	

Statistical distribution of the area average is rather close to the exponential one, as indicated in Fig. 2a without seasonal separation. Overwhelming majority of the wet days represent low precipitation with a steep and monotonous decrease of frequency at higher amounts. Not surprisingly, distribution of the first non-rotated factor loadings (Fig. 2b) is rather similar, indicating that this major component, representing 39 ~64 % of the original variance in the different seasons, are strongly related to the area mean precipitation. Rotation of the factors (Fig. 2c) can not change the situation too much either. The only difference is that here the kurtosis of the distribution is much larger than that of the normal distribution. The point of the matter, i.e. the strong dominance of the low precipitation averages, remains valid for the rotated components, too. On the other hand, the rotation yields more symmetric distribution of loadings and more even distribution of the variance among the chief retained factors, what increases the freedom of clustering.

Seasonal rotated scores are input variables of the cluster analyses, also representing variously distributing fields[12]. Here we present another illustration, based on rotated loadings of the all-year factor analysis of the wet days (6,268 days in the sample). These factors can also be interpreted as sub-regions in which precipitation variations are similar to each other and, at the same time, relatively different from those in the other regions. Fig. 3a indicates the results of this analysis for data of the whole year in 60 stations including Ullung Island. This non-seasonal analysis separates 7 regions with reasonable spatial distribution. The regions are determined by the maxima of the seven rotated loadings at the given station. The majority of stations exhibits nearly 0.7 loading and can be related to one of the seven regions, even if they do not fall into the core of the regions, delimited by the 0.8 loading isolines.

Answer to the fourth question is found in the explained communality, which orders a number to every stations reflecting the proportion of variance statistically explained by linear combination of the retained factors and loadings. If the communality is not close to 1, we establish high proportion of individual variance. For this reason, Ullung Island will be excluded from the country-wide classification of the patterns (Fig. 3b), since only 38 % of its variance is related to the common information represented by the

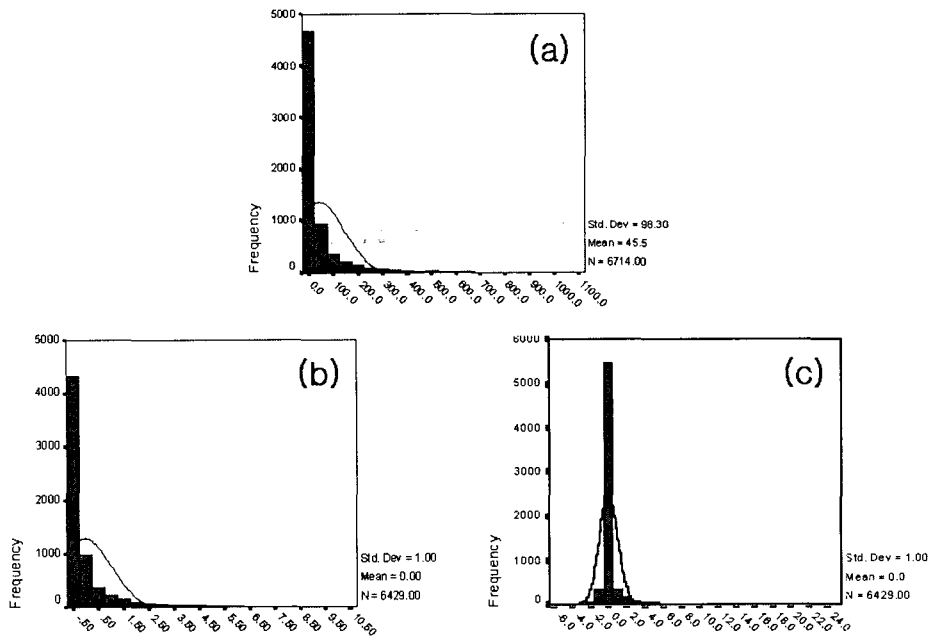

Fig. 2. Distribution histograms of (a) area-mean diurnal precipitation at the 59 stations in 1973-1996 in South Korea, (b) the first diurnal non-rotated factor loadings and (c) the rotated factor loadings. Days being dry at all stations are excluded. Normal distribution is graphically fitted.

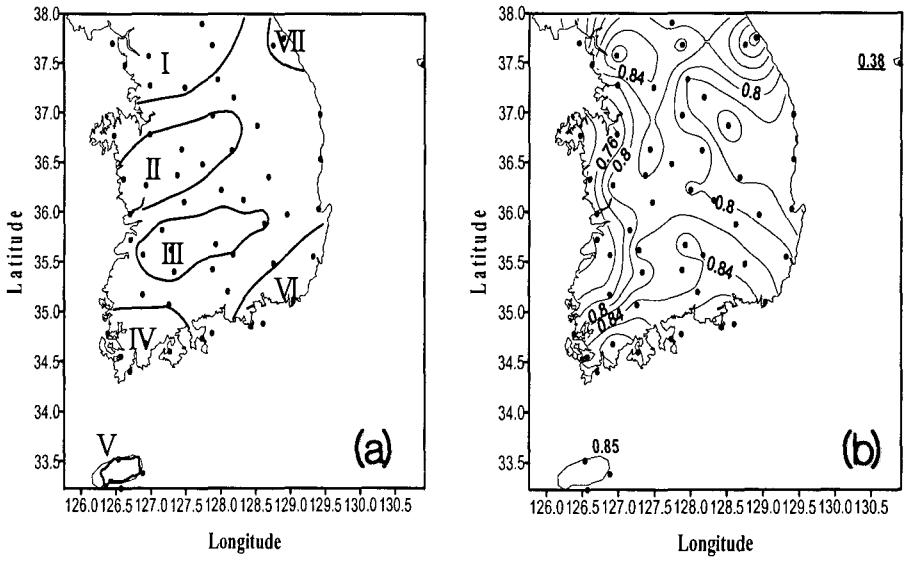

Fig. 3. Distributions of (a) seven regions of precipitation from diurnal factor analysis with no separated seasons, (b) communality represented by the retained factors. Dry days are included.

factors. Jeju Island is worth keeping since the communality is equal or higher here than in the internal part of the peninsula, which might be due to the three stations located on that island.

4.2. Alternative results of cluster analysis

As already mentioned in Section 3.2, function of the between-cluster distance on the decreasing number of clusters does not yield any break, but it represents a smoothly increasing dissimilarity. Figure 4 indicates the behavior of the distances for the last 30 steps in autumn. In its upper module, "Rule 1" indicates the 6 rotated loadings and "95 %" reflects the case of 17 loadings, as bases of the clustering (compare with Table 1). Naturally, the more detailed representation of the diurnal patterns is somewhat more difficult to compress into a fixed number of classes, which is valid also in case of the overall distance (one joint class, i.e. no clustering), too.

In case of the relative classification (pattern correlation, Furthest Neighbors, see lower panel of Fig. 4), however, there are some brakes in the index, which allow to select a natural termination of clustering. This implies that most likely the continuum of the area-mean precipitation is the main reason of the smooth behavior in the Euclidean distance-based way of clustering.

According to the methodology, described in Section

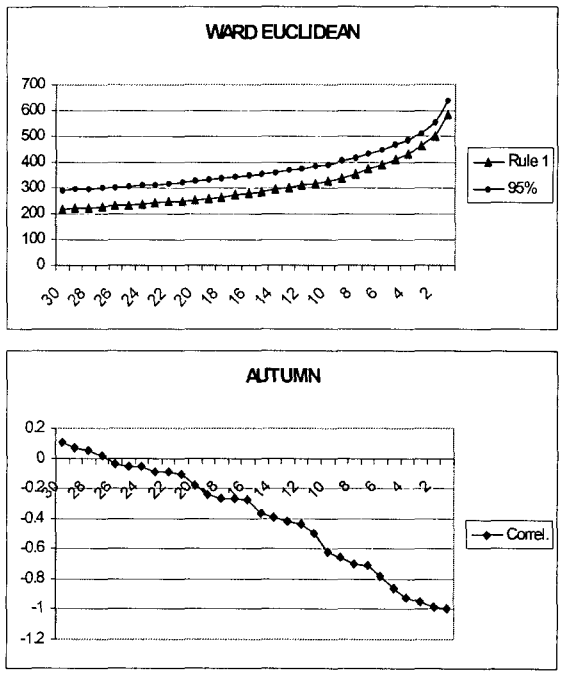

Fig. 4. Curves to support clustering termination: a) Ward method, Euclidean distance smooth increase of the distance index, based on the five factors (according to Rule 1) and 17 factors (explaining 95 % of variance). b) relative classification (pattern correlation and method of Furthest Neighbors) with some breaks in the distance index (i.e. lowest correlation).

3.2, number of clusters and the relative variances of the four candidates and the final selection are comprehended in Table 2. Note that they are related to the wet clusters and cases, without inclusion of the dry days (i.e. cases and clusters characterized by no measurable precipitation at any of the 59 stations). Figures of the Table can be interpreted, as follows:

1) Number of clusters when not any single, non-clustered case remains is rather large and it were difficult to interpret why we had more clusters in spring (29) than in summer (25), also, much more in the transition seasons than in winter. Hence, this candidate should be rejected.

2) Number and seasonal distribution of the $N_{min} \geq 5$ candidate is fairly reasonable and the relative variances are also convincing. Hence, this candidate is worth considering, although the number of clusters is still high.

3) The $N_{min} \geq 1\%$ candidate for clustering is already characterized by convenient numbers of clusters, but the relative variances are not attractive. Especially the summer and autumn values are too big, larger than 50 %.

4) As concerns the last stage before $N_{min} \geq 5$ %, its number of wet clusters is very practical (3-6 clusters), but the unexplained variance is even higher than in the previous case. Hence, although there might be applications, especially, if related to short samples, when the low number of classes are more important than the reduction of variance, it is difficult to recommend a classification for general use, that leaves higher portion of variance unresolved, than explained.

So, there is only one reasonable candidate, the $N_{min} \geq 5$ one, which can be further polished to some extent. It is performed by a systematic search for lower numbers of clusters, where the explained variance is equal or higher. One should note, again, that it is not principally excluded, since the monotonous increase of the distance function parallel to decreasing number of clusters is strictly related only to the reduced number of loadings, not to the original patterns. But, as expected, we found only two possibilities to improve the pre-selected candidates, since the loadings provided fair representation of the patterns. In both cases the number of clusters decreased by one and the efficiency could even be improved by one percent.

4.3. The final classification

The final classification exhibits 14+1; 15+1; 24+1 and 15+1 clusters, in the above-defined winter, spring, summer and autumn periods, respectively. (The+1 cluster indicates the totally dry days with no measurable precipitation at any station.) The cases (days) are distributed very unevenly among the classes, as demonstrated in Figure 5. Having the clusters sorted according to their area-mean precipitation in an increasing order, the frequency of clusters exhibits nearly the opposite distribution. The most frequent wet clusters are characterized by very low amounts of area-mean precipitation in each season. Together with the dry days, the two clusters represent 68, 60, 41 and 66 % of all cases of the seasons starting from winter to autumn, in the above given sequence. In another comparison, the dry days represent 28.4 % of the 24 years, whereas the largest wet clusters cover 30.2 % of that.

In each season there are some clusters with relatively large area-mean precipitation, but their frequency is not high, except for the summer season. The importance of cluster analysis is demonstrated by these high-precipitation clusters: Despite their low general frequency, claiming for equalization by amalgamation into one cluster, strong internal differences of the patterns belonging to one or the other cluster

Table 2. Alternatives of the final classification with the mean relative variance expressed in the non-dry(N-1 cluster) days

Size of the Least Cluster	> 1 member		≥ 5members		Final Selection		≥1%		The Largest Before ≥ 5%	
	Wet Clusters	Expl. Var.(%)	Wet Clusters	Expl. Var.(%)	Wet Clusters	*Expl. Var.(%)	Wet Clusters	Expl. Var.(%)	Wet Clusters	Expl. Var.(%)
Winter	14	see→	14	32	14	32	9	39	4	51
Spring	29	n.a.	16	37	15	36	10	43	5	48
Summer	25	see→	25	43	24	42	12	54	6	63
Autumn	24	n.a.	15	42	15	42	7	57	3	64

* These figures slightly differ from those quoted in Table 3, since the dry days are included.

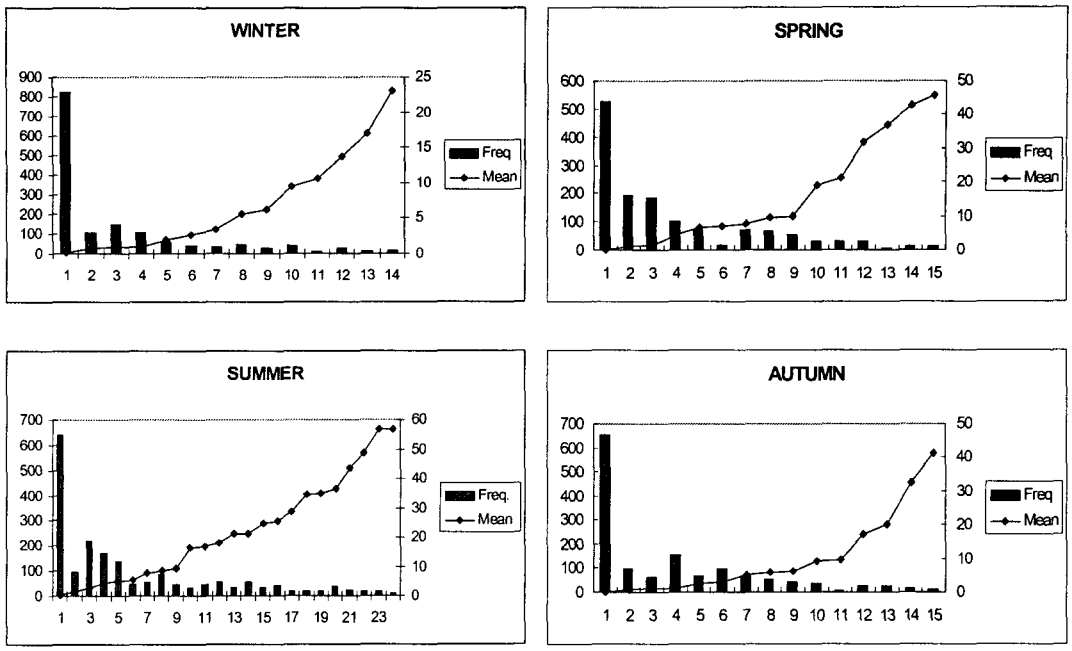

Fig. 5. Frequency and area-mean precipitation of the clusters with non-zero precipitation.

do not allow this unification. In many relations the pattern correlation between the cluster centers is strongly negative. Hence, they should be kept separately.

There is no place to present all the 67 wet clusters, hence we limit ourselves to illustrate the big differences among the clusters. Each season is represented by 4 clusters in Figure 6, according to the following selection: Besides the clusters with the two lowest and the two highest area-mean cluster centers, two other pairs are selected from the middle of the distributions, characterized by strongly negative pattern-correlation, but similar area mean values. The four corresponding pairs are positioned beneath each other with an increasing order of cluster centers, from the left to the right. So, differences of the clusters are recommended to consider mainly between the upper and lower figures, within a given season. The cluster-centers exhibit fairly small-scale patterns of precipitation maxima, which might, however, be related to one or few extreme events, especially in case of small clusters. (See the number of clusters in the headings).

Main statistical characteristics of the final classification are presented in Table 3, incorporating the dry days, as well. Comparing frequencies of the dry cluster to the most frequent wet cluster, the latter ones exhibit higher percentages in winter and summer, but they appear more rarely in the transition seasons. Size of the smallest clusters is always between 5 and 9 members.

The relative variance, explained by the clustering becomes slightly better if we include the dry days into the analysis. Average performance of the classification is as good as 37 %, with better capability in winter and spring (31 and 34 %), but weaker than average in summer and autumn (41 and 40 %). In other words this means that the classification is able to explain the complementary part of variance, i.e. 63 % in average (59~69 % in seasonal extremes).

Speaking more practically, if having been informed about the prevailing cluster in a given day, one substitutes the actual precipitation pattern by the cluster centers, the average squared error of estimation is only 37 (31~41) % of the initial uncertainty, determined just by the knowledge of climatic mean patterns. Relying at a distant analogy, the case of linear regression, in that case similar reduction of uncertainty is achieved by 0.79 (0.77~0.83) correlation coefficients.

Applying weighted relative variances, i.e. dividing the squared deviations by the cluster centers (see Section 3.2), we obtain even better figures. This

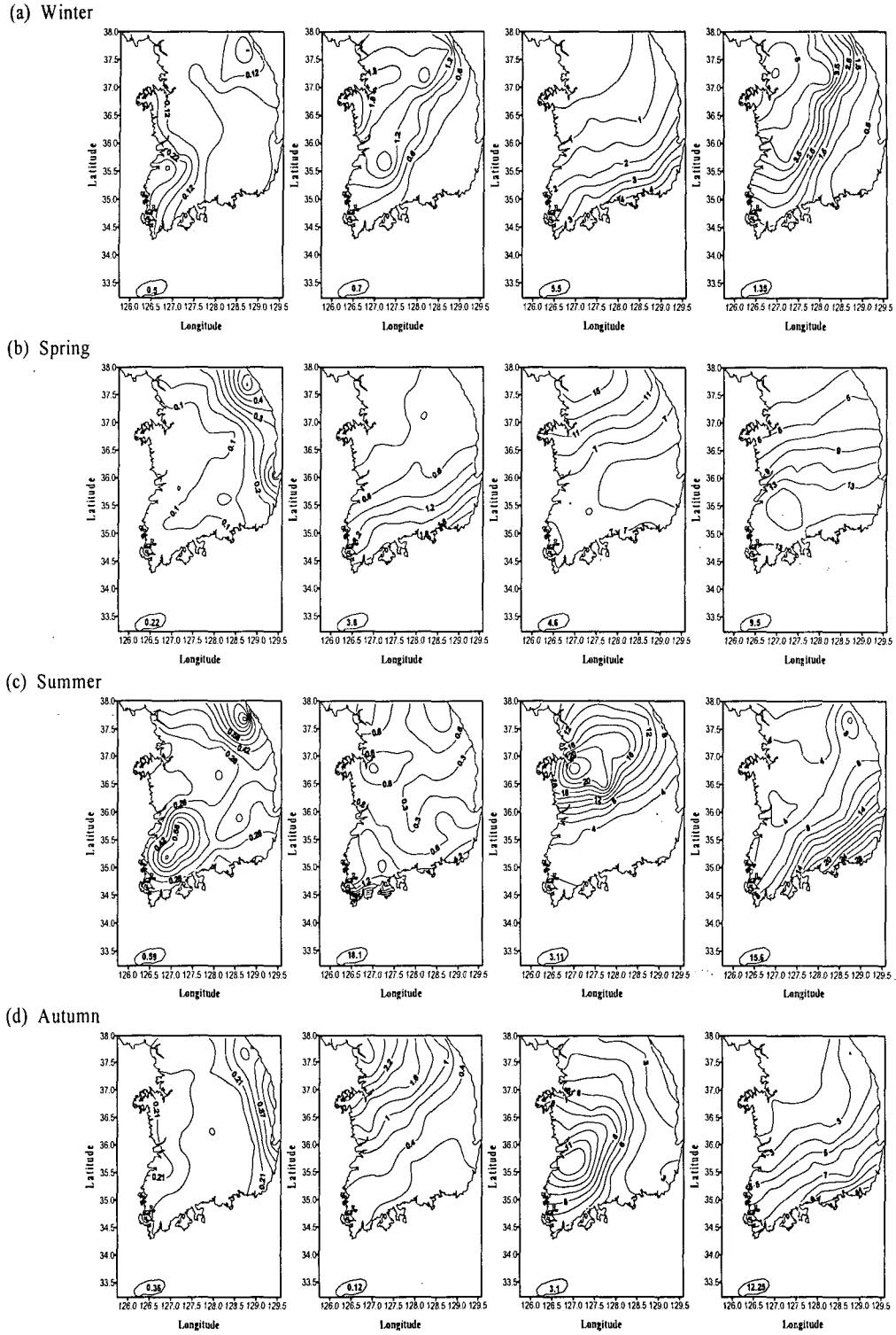

Fig. 6. Cluster centers of four selected maps from each season a) winter, b) spring, c) summer and d) autumn. The four maps of each season represent clusters with the lowest and highest area averages and also pairs with similar, intermediate averages, but strong negative correlation.

Table 3. Main statistical characteristics of the classification. Normalized variance means the proportion of variance to the average, both within a cluster and the whole sample (1973~1996, days with measurable precipitation at least in one of the 59 stations)

Elements / Season	All Days of Season	Wet days	Dry Days	No. of Wet Days in Cluster	Mean Relative Variance(%)	Normalized Variance(%)
Winter (15 Clusters)	2160	1508	652	9-825	31	20
Spring (16 Clusters)	2208	1402	806	5-524	34	25
Summer (25 Clusters)	2205	1952	256	8-640	41	27
Autumn (16 Clusters)	2184	1406	778	5-653	40	17
			Average		37	22

weighted average uncertainty is only 22 % of the unclassified one with a variation between 17 % (autumn) and 27 % (summer).

The last paragraphs demonstrate fairly encouraging numbers, derived from point-wise validation of clustering, which represent spatially averaged gain of information in Korea. To demonstrate information about the spatial variability of performance, we recommend the Figure 7, representing frequency distribution of the remained relative (non-weighted) variance among the 59 stations. The distributions are positively skewed in summer and autumn, i.e. there exist a few stations with poorly explained precipitation, but the majority of them belongs to even lower non-explained variance, than the average numbers of Table 3. The two other seasons are more symmetrical and even the worst stations exhibit slightly above 50 % of non-explained variance.

The standard deviation of the relative variance is only 7-8 % around the mean. So, one can conclude that the cluster analysis, represented by fairly low average percentages of non-explained variances, can also be applied for a small sub-set of stations (including individual ones) with substantial reduction of variance, even in the worst possible cases.

5. DISCUSSION AND APPLICATIONS

This section is devoted to two questions:

1) What alternatives of the performed classfication exist and how they modify the result?

2) How can the obtained classification be applied?

The methodological alternatives include modifications of spatial and time resolution or scope and also technical alternatives to the selected ways of the factor and the cluster analysis.

If considering a smaller area than South Korea, most likely one can obtain lower numbers of clusters with equal efficiency. Such smaller geographical units could be the water basins, but they represent relatively small portion the given area, i.e. too few stations in each watershed. Also, considering the 24 hours averaging time of the precipitation, the larger area represents useful information also for a smaller region, in another part of the country. Hence division of the territory was early rejected.

The given classification is, strictly speaking, a result of the selected 59 stations. However, the final result is likely resistant against some increase or decrease in density of the stations, since we applied a preliminary factor analysis to focus on the most relevant common information and to study the spatial correlation. To check this dependence on the number of stations might be crucial when an extension of the classification to the distant past (fewer stations) is considered.

If considering, the extension of the classes in space, i.e. for the whole Korean Peninsula, than the present work has to be performed from the beginning. It could also be of scientific interest, how the clusters of the larger area coincide with the ones, derived just at a part of it.

A specific example of spatial extension, the inclusion of Ullung Island would likely cause larger number of clusters, due to its large individual variability, as demonstrated in Section 4.1.

Alternatives in time resolution could strongly mod-

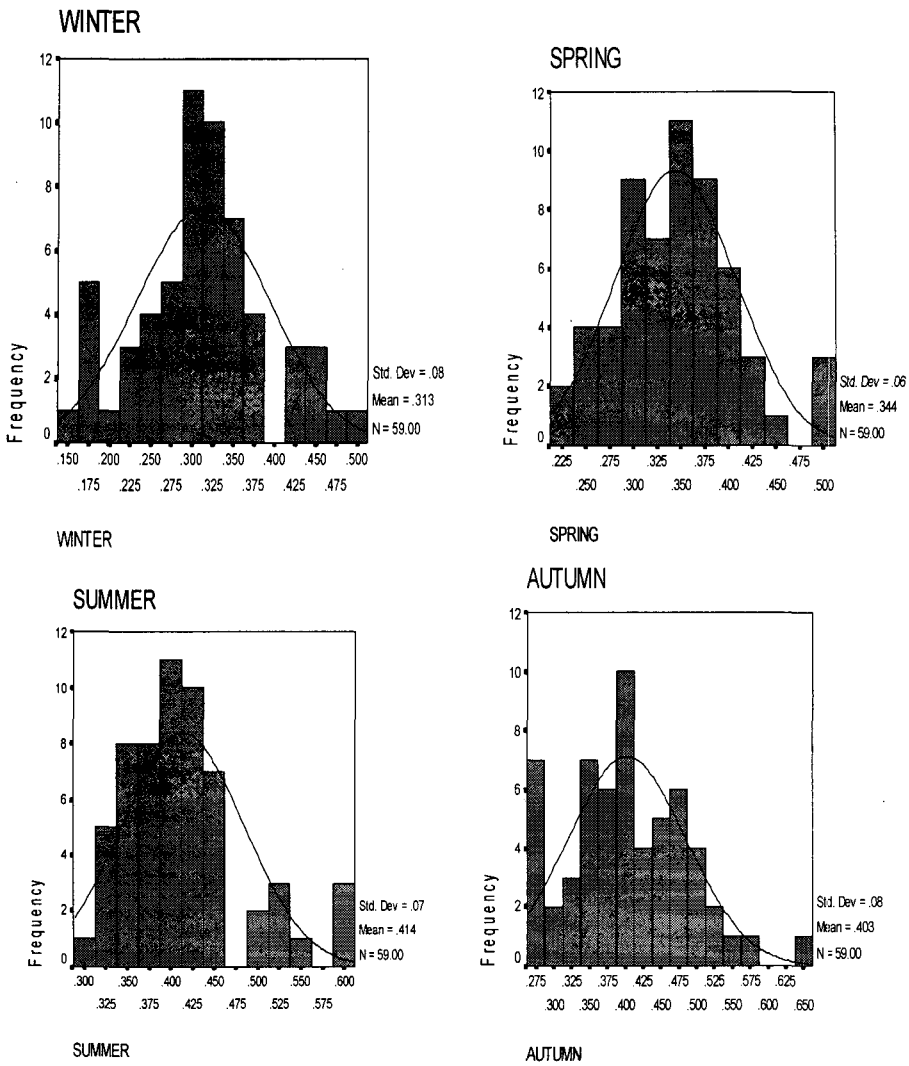

Fig. 7. Frequency distribution of the non-explained relative variance among the 59 stations. (Normal distribution is graphically fitted.)

ify the picture, especially if dividing the summer season into more parts. However, our trial, delimiting 9 sections of the year with non-equal duration, following a very detailed analysis of Fig. 1, yielded too small numbers of wet days to find enough clusters to represent the strong variability of precipitation patterns. On the other hand, the recommended seasonal separation is quite natural and it may cause considerable inhomogeneity of the area-mean precipitation only in the autumn season (Sept. 15 - Dec. 15).

The other aspect of time resolution is the 24-hour averaging. Since precipitation is mainly connected to meso-scale objects of shorter, than one day life-time or, at least large spatial shifts within this period, the averaging time strongly influences the classification. Although the day is a natural resolution in the majority of practical respects, where the diurnal classification can be applied (see below), it is recommended to repeat the classification for other time resolutions, too.

Considering the performed factor analysis, we already mentioned the parallel classification, performed by applying larger number of retained factors, explaining more spatial details, i.e. 95 % of variance. Although, it was also demonstrated that the cluster distances were higher, it could well be the case that

the situation became the opposite, if considering the real patterns, better resolved in the first step. However, the retained further spatial details caused enhanced longevity of isolated cases in winter and in summer. They diminished at later stages of cluster analysis, yielding 10 and 15 clusters according to the above used first criteria for selection with near and above 50 % of non-explained variance. If adjusting the two other seasons to the number of clusters in summer, we could obtain a classification with 10, 11, 15 and 10 clusters from winter to autumn, but much higher 47 (42~52) % non-explained variance, than in the recommended final classification. Of course the situation could be improved by retaining single (non-classified) days, but in this case, we left dangerous extreme events out of the scope of classification.

The cluster analysis could alternatively be performed in three respects. First, all point-wise precipitation patterns could be transformed, even before the factor analysis, in order to ensure less skewed distribution. Although, when we prepared the cluster analysis after logarithmic transformation (assigning 1 values to stations and days with no precipitation) the result was quite unpleasant: One cluster contained almost every day with single cases in the rest of the clusters. This "snowball-effect" completely hinders the application of this transformation. Other solutions, e.g. square root transformation, would be less acceptable from both physical and practical point of view. That is, they were fixed to the given non-physical transformation, preferable in a limited (if any) circle of applications. It is hard to imagine, either, that a clustering optimized for a given set of transformed values, would perform better for the original diurnal precipitation.

The second aspect could be the use of other than Euclidean distance (besides the pattern correlation, discussed in Section 4.2). Some alternatives were also tried, in this respects, but they gave either fairly similar results to the recommended one, or they led to the above snowball effect due to suppression of real differences among the patterns. (Some versions were not even tried, following the principle of not distorting natural precipitation units or comparisons.)

There are also alternative rules of cluster amalgamation to the Ward's one. But, method of Furthest Neighbors, for example, led to the snowball effect again, whereas other versions, operating with the cluster centers, gave quite similar results to the recommended one. According to its definition, the Ward's method ensures the lowest non-explained within-cluster variance in the space of involved factor loadings, which, in turn, promise good chance to provide a similar optimum feature in space of the precipitation patterns, as well.

In the recommended classification, the 15~26 clusters are near the upper margin of acceptance. At the present technological level, it is not determined by the amount of computations but by the extensive need for duration of the series one wishes to relate to the calendar of clusters. In Section 4.3 we could see that some precipitation clusters occurred just a few times in the 24 years long basic period. This number of cases (59) does not allow us to provide any statistical estimation. On the other hand, frequency of the majority of clusters is substantially larger, i.e. the lower statistical moments can already be estimated with confidence.

The overall quality of the classification, i.e. the information surplus obtained by the knowledge of the actual cluster is fairly high: It allows us to reduce the uncertainty by 63 %, in average. So, besides the final classification recommended for wider use, there is some possibility to find another ad-hoc compromise, by joining several clusters despite the decreasing quality of separation, that is designed to given limitations of the application. As indicated in Section 4.2 if one allows us to explain 40~50 % of the variance, the number of clusters can be as low as 10 or even less.

Technical aspects of the classification imply the calendar and the cluster-statistics to be obtained; the required dimensions of the related data and possibilities to extend the calendar before or after the 1973~96 period.

In many aspects it is enough to apply just the calendar of the types and the related fields can be determined independently. Of course, for meaningful conclusions on these relations, careful consideration of the cluster-means and, sometimes, of the variances is also required. For other purposes, the field characteristics of the classes should be used in an operative basis, to compare them with the on-going or forecasted diurnal events.

It is advisable to have the related physical, chem-

ical or ecological characteristics also in the form of area patterns for Korea. These patterns may be available in much lower resolution than the original 59 stations. Even if there is one single station or a particular region of Korea in the focus of the interest, it may be useful to apply the classes (instead of point-wise data, or regional averages) in some cases. The classes can apparently represent implicit information on the related meteorological or environmental variables that are not available locally.

It is also of practical importance, how we can apply the elaborated classification for precipitation patterns that are not involved in the 1973~1996 basic period. If the number of stations is the same 59, which relates mainly to the more recent years, one can simply select the most similar cluster center for each day, however in a part of cases the within-cluster standard deviations should also be considered. Since the distribution of precipitation is far from normal, a more established statistical decision-making procedure might be required for more exact solution of this post-classification. An alternative of this more sophisticated algorithm can be the application of the non-hierarchical, K-means clustering, with the recommended cluster centers to start with. In this process, however some correction of the final number, really represented by the additional years, could be needed, as some rare clusters may not occur in the short period.

If just a few stations are missing (the near-1973 years) similar procedure can be recommended by using the common stations. A more sophisticated elaboration could be achieved by a regression-based computation of the factor loadings that are related to the basic 5-8 factors in the given seasons. This can be performed just for the existing stations, but the performance of the K-means clustering deals with the loadings, starting from the recommended cluster means, anyway.

If more stations are missing, especially from regions, where one or more cluster-centers exhibit high values of precipitation, then these patterns can not be sorted to the basic classification due to lack of information. If there is a long period with smaller number of stations, which allows to do so, a new cluster analysis is recommended, repeating all steps of the work, i.e.:

1) Factor analysis of the existing stations and selection of the number of factors to retain.
2) Rotation of the factors and registration of diurnal factor loadings for further use.
3) Cluster analysis of the days, represented by non-correlated information of the loadings.
4) Statistical post-processing of the results to obtain natural precipitation cluster statistics.

The aim of the factor analysis is not to reduce the matrices before cluster analysis, but to consider spatial correlation of precipitation fields. Hence, it is not recommended to omit the first two steps, even in case of small number of stations. Otherwise the distance function should consider this correlation, but it is not a default option in many statistical software products.

Acknowledgements

This research was performed for as a part of "A Study on the Typhoon Monitoring and Prediction System Development" supported by the Meteorological Research Institute (METRI) in the Korean Meteorological Administration. Contribution of Mr. Oh H.-Y. by drawing the maps is deeply appreciated. The STATISTICA for Windows (10.0) software was used for factor-analysis and cluster-analysis computations.

REFERENCES

1) Seo, A. S. and C. H. Joung, 1982, The climatic factor analysis of precipitation temperature and sea-level pressure over Korea using empirical orthogonal functions, J. Korean Meteor. Soc., 26, 40-50.
2) Ho, C. H. and I. S. Kang, 1988, The variability of precipitation in Korea, J. Korean Meteor. Soc. 24, 38-48 (in Korean).
3) Moon, Y. S., 1990, Division of precipitation regions in Korea through cluster analysis, J. Korean Meteor. Soc., 26, 203-215.
4) Park, J. G. and S. M. Lee, 1993, A regionalization of annual precipitation over South Korea, J. Korean Meteor. Soc., 29, 117-125.
5) Lee, D. K. and J. G. Park, 1999, Regionalization of summer rainfall in South Korea using cluster analysis, J. Korean Meteor. Soc., 35, N4, 511-518.
6) Hair, J. F., R. E. Anderson, R. L. Tatham and W. C. Black, 1998, Multivariate data analysis (Fifth Ed.), Prentice Hall, New Jersey, 730pp.

7) Storch, H. and F. W. Zwiers, 1999, Statistical analysis in climate research, Cambridge Univ. Press, Cambridge, UK, 484pp.
8) Bartzokas, A. and D. A. Metaxas, 1993, Covariability and climatic changes of the lower troposphere temperatures over the Northern Hemisphere, Il Nouvo Cimen., 16C, 359-373.
9) Jolliffe, I. T., 1993, Principal Component Analysis: A beginner's guide - II Pitfalls, myths and extensions, Weather, 48, 246-253.
10) Anderberg, M. R., 1973, Cluster analysis for Applications, Academic Press, New York.
11) Ward, J. H., 1963, Hierarchical grouping to optimize an objective function, Journal of American Statistical Association, 58, 236pp.
12) Mika, J., B. J. Kim, C. H. Cho and Yong-Sang Kim, 2001, On the Empirical Eigen-Structure of Daily Precipitation in Korea, Korea J. Atmospheric Sciences, 4(2), 105-116.