Developments in Statistics
Andrej Mrvar and Anuška Ferligoj (Editors)
Metodološki zvezki, 17, Ljubljana: FDV, 2002

The Use of EOF Analysis for Preparing the Phenological and Climatological Data for Statistical Downscaling - Case Study: The Beginning of Flowering of the Dandelion (*Taraxacum officinale*) in Slovenia

Klemen Bergant, Lučka Kajfež-Bogataj, and Zalika Črepinšek[1]

Abstract

Phenological observations are a valuable source of information for investigating the relationship between climate variation and plant development. Potential climate change in the future will shift the occurrence of phenological phases. Information about future climate conditions is needed in order to estimate this shift. General Circulation Models (GCMs) provide the best information about the future climate change. They are able to simulate reliably the most important mean climate features in large-scale, but they fail in regional scale because of their low spatial resolution (about 2.5°). Researchers dealing with impact studies need, as an input, estimation of potential climate change on a regional or even a local level. A common approach to bridge the scale gap between the large and regional, or local-scale is statistical downscaling. The basic idea behind statistical downscaling is to use the observed relationships between the large-scale climate parameter (predictor) and local-scale climate or climate dependent parameter (predictand), for the projection of GCM results on a regional or local-scale. In our case, beginning of flowering (BF) of dandelion (*Taraxacum officinale*) on 21 locations in Slovenia was related to average January (TJ), February (TF) and March air temperature (TM) on the 1000-mb pressure level across Central Europe. Statistical models were developed and tested with data for the time period 1958-1999. Before developing statistical models, Empirical Orthogonal Functions (EOF) Analysis was performed on climatological and phenological data. The EOF Analysis was used for outlier detection in the data, and for the reduction of input (predictor) data in the developed statistical models. Multiple linear

[1] Department of Agronomy, Biotechnical faculty, Jamnikarjeva 101, SI-1000 Ljubljana.

regression (MLR) models were used to relate the BF with expansion coefficients of first three EOFs for TJ, TF and TM. Criteria for testing the quality of statistical models were variability of predictand data explained by the models, residual analysis and cross-validation. The models explain on average 68% of predictand variability. The correlation coefficient between observations and cross-validation estimations is on average 0.71. Absolute values of studentized residuals are with few exceptions lower than the critical value (2.75 for $\alpha = 0.01$) and have no typical patterns. None of the unusual residuals represent an influential point - Cook's distance does not exceed the value of 0.43. The results show a strong relationship between predictand BF and predictors TJ, TF and TM. Developed models can be used for the estimation of missing BF values in historical records or to downscale the GCM results in regional climate change studies.

1 Introduction

Phenological observations are a valuable source of information for investigating the relationship between climate variation and plant development (Spano et al., 1999). Potential climate change in future will shift the occurrence of some phenological phases of plants. Information about future climate is needed in order to estimate this shift. On large-scale, General Circulation Models (GCMs) are able to simulate reliably the most important mean features of the global climate (Zorita and von Storch, 1999). The GCMs are based on mathematical formulation of climate system dynamics and are run to simulate past and future development of a global climate system. Unfortunately, they have one main disadvantage - low spatial resolution. The typical spatial resolution of GCMs is about 2.5°. Their results are representative in a large, but not in a regional or even a local-scale. Researchers dealing with impact studies usually need, as an input, estimation of potential climate change on a regional or on a local level. For the estimations on the local level, it is necessary to bridge the gap between the large-scale, for which we can use GCM results, and the local-scale.

A common approach to bridge the scale gap is statistical downscaling (Rummukainen, 1997; Wilby and Wigley, 1997; Zorita and von Storch, 1999). The basic idea of the statistical downscaling is (1) to look for the relationships between the large-scale climate parameter (predictor) and the local-scale climate or climate dependent parameter (predictand) in the observed data, and (2) to use these relations for the projection of GCM simulation results for past, present or future climate on the regional or local-scale. The paper treats only the first part of the statistical downscaling procedure.

The beginning of flowering of dandelion (*Taraxacum officinale*) at 21 locations in Slovenia was related to the average January, February and March air temperature on the 1000-mb pressure level across the Central Europe. Temperature data and beginning of flowering dates for the years between 1958 and 1999 were used for development and testing the statistical models. Before developing the statistical models, EOF analysis was performed on climatological and phenological data.

The aim of our study was to evaluate the role of EOF Analysis in the data preparation phase of statistical downscaling process, and to evaluate the ability of developed statistical models to estimate the unknown predictand values on the base of known predictor values.

2 Data and methods

2.1 Phenological and climatological data

The Dandelion (*Taraxacum officinale*) is important as medical, food and apicultural plant. The reported life zone of the dandelion has an average annual air temperature of 5-26°C and an annual precipitation of 300-2700 mm (Simon et al., 1984). Consistent with its life zone, dandelion is found almost on entire area of Slovenia, except in high altitudes, where the average annual air temperature drops below critical value. Due to its spatial extension and easy detectable phenophase, the observations of the beginning of flowering times for the dandelion in Slovenia are numerous and have good quality.

The time series of the beginning of flowering dates (BF) of the dandelion for the period 1958-1999 for 21 sites in Slovenia (Table 1) were used as the predictand data in our study. The phenological data were collected and verified by Environmental Agency of the Republic of Slovenia. Average occurrences of BF phenophase across the years 1958-1999 for single phenological stations are presented in Table 1, together with their geographical data.

In general, spring flower phases are strongly influenced by the air temperature of the previous months (Maak and von Storch, 1997). Because the beginning of flowering of the dandelion in Slovenia starts on average from the end of March to the beginning of May, the average monthly air temperatures for January (TJ), February (TF), and March (TM) were chosen as predictors. Free atmospheric data (air temperature on 1000-mb pressure level) instead of surface data (near ground air temperature) were chosen as predictors, because GCMs are not capable of simulating the surface climate very well (Rummukainen, 1997). The air temperature data were taken from the dataset of common NCEP (National Centers for Environmental Prediction, USA) and NCAR (National Center for Atmospheric Research, USA) project. The NCEP/NCAR reanalysis data (Kalnay et al., 1996) for TJ, TF and TM with spatial resolution of 2.5° for the area with longitudes from 2.5°E to 22.5°E and latitudes from 35°N to 55°N, including 81 grid points across the Central Europe, were related to beginning of flowering of dandelion at 21 sites in Slovenia by multiple linear regression models. The size of the predictor area corresponds to criteria of skillful scale of GCMs (von Storch et al., 1993), which should include at least eight GCM grid points. At the lowest end of the spatial resolution of GCMs, they have little or no skill (von Storch, 1995).

Table 1: Selected phenological stations, their geographical data and average occurrence of beginning of flowering of dandelion (BF) in the time period 1958-1999.

Station	Site	latitude [°]	longitude [°]	altitude [m]	BF [Julian days]
1	PODLIPJE	46°38'	15°10'	760	108.5
2	MARIBOR-TEZNO	46°32'	15°39'	275	100.9
3	GORNJA RADGONA	46°40'	16°00'	205	94.2
4	STARŠE	46°28'	15°46'	240	97.5
5	RATEČE-PLANICA	46°30'	13°43'	864	121.2
6	RADOVLJICA	46°21'	14°11'	495	107.8
7	SORICA	46°13'	14°02'	820	124.1
8	MOZIRJE	46°20'	14°58'	347	102.1
9	CELJE	46°15'	15°15'	244	101.5
10	SLOVENSKE KONJICE	46°20'	15°26'	332	101.5
11	KLENIK PRI VAČAH	46°07'	14°51'	550	103.2
12	LJUBLJANA BEŽIGRAD	46°04'	14°31'	299	103.1
13	MOKRONOG	45°57'	15°09'	251	101.0
14	NOVO MESTO	45°48'	15°11'	220	95.3
15	ČRNOMELJ-DOBLIČE	45°34'	15°09'	157	95.7
16	KOČEVJE	45°38'	14°52'	461	107.5
17	POSTOJNA	45°46'	14°12'	533	107.0
18	TRENTA	46°24'	13°45'	748	120.1
19	SLAP PRI VIPAVI	45°50'	13°56'	137	89.8
20	ŠMARJE SAP	45°58'	14°37'	342	93.5
21	ILIRSKA BISTRICA	45°34'	14°15'	414	93.6

Table 2: Summary of data used for developing statistical downscaling models.

Predictor data

- *Average monthly air temperature on 1000-mb level*
  - 81 grid points homogeneously spread across Central Europe (see Fig. 2) (I = 81)
    (longitudes: 2.5°E - 22.5°E; latitudes: 35°N - 55°N; spatial resolution of data: 2.5°)
  - Time series for January, February, and March for single grid points (TJ, TF, TM)
  - Time period: 1958-1999 (42 years) (T = 42)
  - Origin: NCEP/NCAR Reanalysis project

Predictand

- *Beginning of flowering dates of dandelion* (*Taraxacum officinale*)
  - 21 sites in Slovenia (heterogeneously spread across country – see Table 1) (J = 21)
  - Time series of beginning of flowering dates (in Julian days) for single sites (BF)
  - Time period: 1958-1999 (T = 42)
  - Origin: Environmental Agency of the Republic of Slovenia

Before developing the statistical models relating the beginning of flowering of dandelion with the air temperature of the previous months, the Empirical Orthogonal Function Analysis (Björnsson and Venegas, 1997; Wackernagel, 1998;

von Storch and Zwiers, 1999) was employed on predictor (TJ, TF and TM) and predictand data (BF).

2.2 Empirical orthogonal functions analysis

In climatology, Principal Component Analysis (PCA) has been transposed from a multi-variate into a multi-station space-time context, where the technique received the name of Empirical Orthogonal Functions (EOF) Analysis (Wackernagel, 1998). In general, spatial data can be written as:

$$S_{i,t} = \sum_{k=1}^{K} a_{t,k} EOF_{k,i} + n_{i,t} \qquad \begin{array}{ll} t = 1,\dots,T & T \text{ - number of observations (years)} \\ i = 1,\dots,I & I \text{ - number of grid points} \\ k = 1,\dots,K & K \text{ - number of retained EOFs} \end{array} \tag{1}$$

In equation (1) i is a grid point index, t is the time index, EOF_k is the kth EOF spatial pattern, and $a_{t,k}$ is the expansion coefficient of this pattern at time t. The $n_{i,t}$ is the part of time variability of S on grid point i not described by the leading K patterns at time t, and is assumed to be small. EOF patterns are the eigenvectors of the covariance matrix **R**. If the anomalies of the predictor S at different times t and on different grid points i are written in matrix **S**, $S_{i,t}$ represent its elements, i.e. $\mathbf{S} = \left\| S_{i,t} \right\|$; and the covariance matrix can be estimated by the equation (2) (Björnsson and Venegas, 1997).

$$\mathbf{R} = \frac{1}{T-1} \mathbf{S} \cdot \mathbf{S}' \tag{2}$$

In equation (2), T represents the number of predictor data at different times t on each of I grid points. Now, EOF patterns can be estimated as eigenvectors of covariance matrix (equation 3).

$$\mathbf{R} \cdot \mathbf{E} = \mathbf{\Lambda} \cdot \mathbf{E} \tag{3}$$

In equation (3) the matrix **E** contains the EOF patterns, i.e. $\mathbf{E} = \left\| EOF_{k,i} \right\|$ and matrix Λ contains the eigenvalues λ_k of the covariance matrix on its diagonal.

The EOF analyses were employed to phenological and climatological data from two aspects:

- data compression;
- outlier detection.

In the first case the EOF analysis was used for reducing the number of input data for further analyses, considering the fact that the first few EOFs explain most of the variability of data (Wackernagel, 1998; von Storch and Zwiers, 1999). Further the results of the EOF analysis were also used for detecting the possible outliers in the climatological and phenological data.

2.3 Statistical downscaling

2.3.1 Development of the statistical model

As mentioned before, the statistical downscaling approach was used to bridge the gap between the large-scale and the local-scale. The observed anomalies of BF data ($\vec{B}_a$) with regard to average values for time period 1958-1999 for a single site were related to EOF expansion coefficients of TJ, TF and TM, by multiple linear regression models (Krzanowski, 1998).

$$\vec{B}_a = \mathbf{A} \cdot \vec{L} + \vec{\varepsilon} \tag{4}$$

The columns of matrix **A** in equation (4) are the time series of the EOF expansion coefficients for TJ, TF and TM with elements $\left\| a_{t,k}^{TJ} ; a_{t,k}^{TF} ; a_{t,k}^{TM} \right\|$. Vector $\vec{\varepsilon}$ represents the part of variability of BF data not explained by linear model, and vector $\vec{L}$ represents the corresponding regression coefficients, which can be estimated by using equation (5) (Krzanowski, 1998).

$$\hat{\vec{L}} = \left(\mathbf{A}' \mathbf{A}\right)^{-1} \cdot \mathbf{A}' \vec{B}_a \tag{5}$$

A similar approach was used by Schubert (1998) for downscaling the local extreme temperature changes in Australia, and by Benestad (1999a, 1999b) for predicting the monthly mean land surface temperatures in Norway.

2.3.2 Testing the statistical model

Criteria for testing the quality of statistical models were the following:

- variability of predictand data explained by the models;
- residual analysis;
- cross-validation.

In the case of explained variability, the coefficient of multiple determination was used as a measure of model quality (von Storch and Zwiers, 1999); and in the case of the residual analysis, studentized residuals and Cook's distance (Krzanowski, 1998) were used.

In the cross-validation approach (Benestad, 1998; von Storch and Zwiers, 1999), the data for a single observation time are excluded in the construction of statistical models. Constructed models are subsequently used to estimate the value of the predictand that was excluded. The procedure is repeated T times, where T is the total number of observations- in our case years (T = 42). As a measure of the ability of the statistical models to predict the unknown predictand values, Pearson's correlation coefficient (Krzanowski, 1998) between observed and estimated values was chosen.

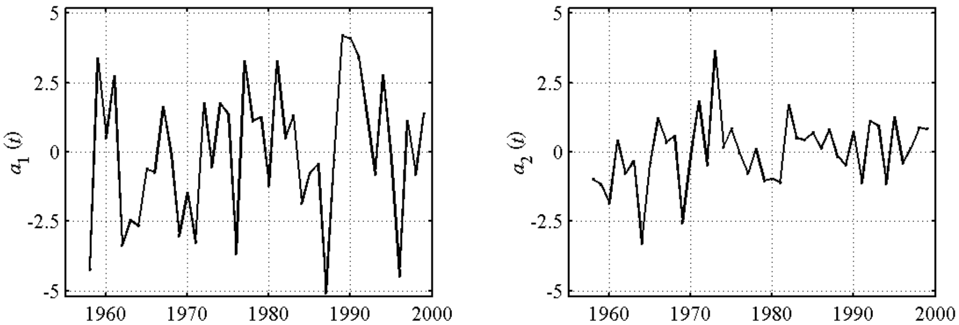
3 Results

The first aspect of the EOF analysis in our case was to reduce the number of predictor data. The time series of TJ, TF and TM data for all 81 grid points (243 time series) can not be used directly in statistical models relating the beginning of flowering dates of dandelion with the air temperature across the Central Europe. The number of time series is too large and the time series are not independent to each other. The EOF Analysis enables the transformation to time series of the amplitudes of typical orthogonal spatial patterns (*EOF*s), where the first few pairs of amplitudes and patterns cover most of the space-time variability in predictor data. The entire predictor dataset for 3 month (M), 81 grid points (I) and 42 years (T), contains 10206 ($M{\cdot}T{\cdot}I$) temperature data. In case of previous employment of the EOF analysis and considering the fact, that usually first 3 EOFs (K) with their expansion coefficients explain most of the variability of data (Table 3), we reduce the number of predictor data on 1107 ($M{\cdot}[K{\cdot}T+K{\cdot}I]$) which is less than 11% of original data. In later statistical analysis only time series of expansion coefficients are used. This further reduces the number of predictor data included in statistical models on 378 ($M{\cdot}[K{\cdot}T]$), which is less than 4% of original data. From 243 ($M{\cdot}I$) time series, the predictor data is reduced to 9 ($M{\cdot}K$) time series. As an example, first three EOFs and corresponding expansion coefficients for TM are presented on Figs. 1 and 2. Combining the three spatial patterns (Fig. 1) and their time amplitudes (Fig. 2) through the equation (1), most of the spatial and time variability of selected TM data, can be explained.

Table 3: Percentage of explained variability of average monthly air temperature data on 1000-mb pressure level (January - TJ, February - TF and March - TM) and beginning of flowering of dandelion data (BF) with first three EOFs.

EOF R^2 [%]	TJ	TF	TM	BF
1	74.7	71.2	75.9	75.9
2	11.5	16.2	11.3	4.7
3	5.5	6.5	6.5	2.9
Sum	91.7	93.9	93.1	83.5

The second aspect of the EOF analysis was to find outliers in data. EOF Analysis was performed on phenological data (beginning of flowering of dandelion - BF). With rotation of 3D graph of first three EOFs expansion coefficients no significant outlier was found (Fig. 3).



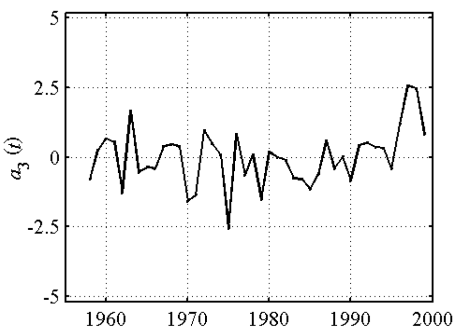
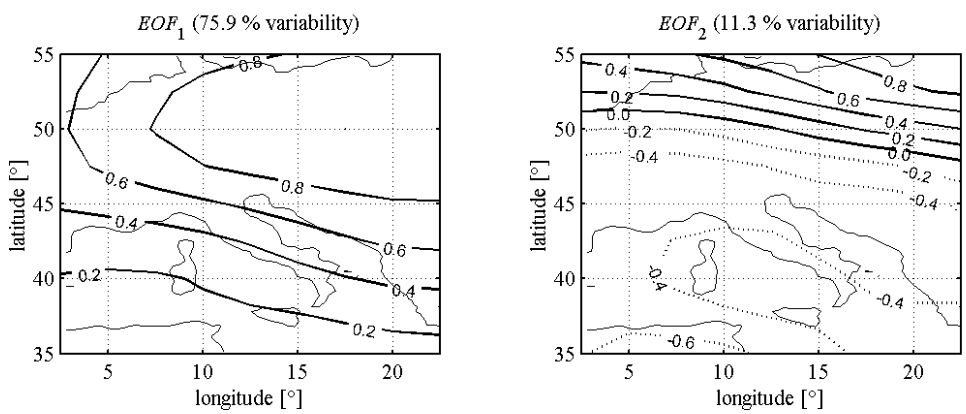


Figure 1: Expansion coefficients of first three EOFs of average March air temperature on 1000-mb level for time period 1958-1999.



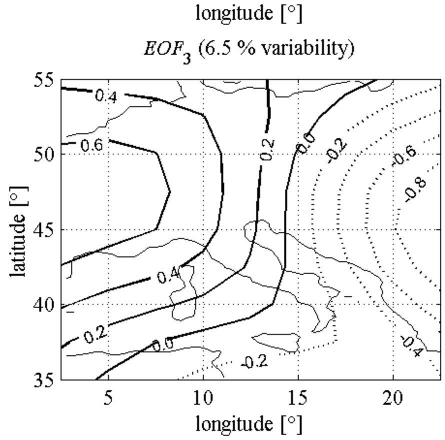


Figure 2: First three EOFS of average March air temperature on 1000-mb level for time period 1958-1999.

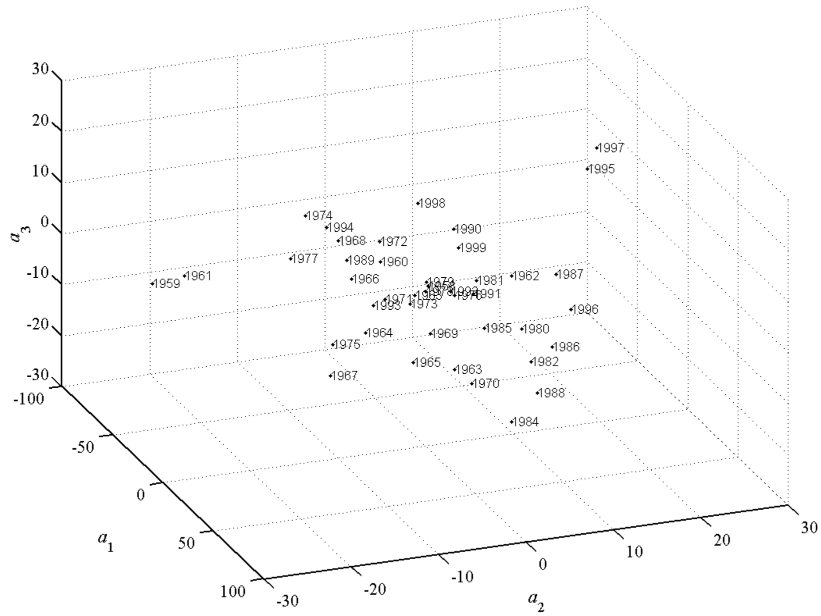


Figure 3: 3D graph of expansion coefficients of first three EOFs of beginning of flowering dates of dandelion for time period 1958-1999 for 21 sites in Slovenia.

The EOF analysis of outliers was performed also on predictor data (TJ, TF and TM), and no significant outliers were found either.

After input data reduction and detection of possible outliers, statistical models were developed for single sites. Expansion coefficients of the first three EOFs of TJ, TF and TM were retained and used in multiple linear regression models. Developed models explained a great part of the time variability of the predictand data. The explained variability differs from station to station in the range between 57% and 84%. On average, models explain 68% of the predictand data variability (Table 4).

Table 4: Explained variability (R^2) of BF data with statistical models for every 21 phenological stations

Station	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
$\boldsymbol{R^2}$ **[%]**	63	81	69	67	61	61	63	72	78	75	84	71	67	74	72	64	74	57	60	78	80

The correlation coefficients between observed and estimated values, produced by cross-validation, take values from 0.54 to 0.87, on average 0.71 (Table 5). They are statistically significant at 99% confidence level with critical value 0.39 (Košmelj and Kastelec, 1996) for all stations.

Table 5: Correlation coefficients (r) between observations and cross-validation BF data for every 21 phenological stations.

Station	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
r×100	66	83	69	68	62	62	66	74	79	76	87	73	69	76	75	65	74	54	60	79	80

For all sites, the average absolute values of the residuals are lower than 10^{-6}, and the standard deviation vary from 3.9 to 7.6 days (Table 6).

Table 6: Standard deviation (s) of studentized residuals for statistical models for every 21 phenological stations.

Station	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
s [days]	7.3	3.9	6.1	6.2	5.8	5.8	5.3	5.4	3.9	5.3	3.8	4.8	5.2	5.9	5.5	7.6	5.9	6.6	6.5	4.5	5.3

Studentized residuals are with few exceptions (1 unusual residual for 10 sites, 2 unusual residuals for 4 sites and 3 unusual residual for 3 sites) in the range between -2.75 and +2.75, which is the critical $t(n\text{-}p\text{-}1)$-value ($n = T = 42$, $p = M{\cdot}K = 9$) in case of $\alpha = 0.01$. They have no typical pattern (not shown). None of the unusual residuals has Cook's distance larger than 1 (values are lower than 0.43), so they do not influence significant on the developed regression models.

4 Conclusions

The Empirical Orthogonal Functions Analysis significantly reduces the large sets of input data needed for statistical downscaling. From 243 time series of average monthly air temperature on 1000-mb pressure level for single grid points for the first three months, the input data were reduced to nine time series of the expansion coefficients of the leading EOFs. The large number of predictor time series is reduced to such extend, that can be used in a simple linear regression models relating the beginning of flowering of the dandelion with the air temperature data.

The Empirical Orthogonal Functions Analysis is a useful tool in outlier detection of available predictand and predictor data, since the outliers could influence the quality of developed models. The outlier can be detected on a 3D graph of the expansion coefficients of the first three EOFs as a point isolated from the others. In our case no such point was found.

In general, the Empirical Orthogonal Functions Analysis can be found as an important and common tool in the phase of predictor and predictand data preparation for the development of statistical downscaling models.

The developed statistical models indicate a strong relationship between the Central European air temperature on 1000-mb level in first three months of the year and the beginning of flowering dates of the dandelion in Slovenia. The models explain a great part of the variability of predictand data. A part of the unexplained variability could probably be related to some other climatological parameters, for example precipitation (e.g., Spano et al., 1999).

The results of cross-validation indicate the ability of developed statistical models for estimating the unknown values, which was one of our main purposes to test. In conjunction with residual analysis, we can conclude that the models can be used for the estimation of missing phenological data in historical records on the base of known large-scale temperature data, or for the projection of General Circulation Model simulation results to the occurrence of phenophase in different climate conditions.

References

[1] Benestad, R.E. (1998): CCA aplied to statistical downscaling for prediction of monthly mean land surface temperatures: model documentation. *DNMI Report.* **28**.

[2] Benestad, R.E. (1999a): MVR aplied to statistical downscaling for prediction of monthly mean land surface temperatures: model documentation. *DNMI Report*, **2**.

[3] Benestad, R.E. (1999b): Pilot studies on enhanced greenhause gas scenarios for Norwegian temperature and precipitation from empirical downscaling. *DNMI report*, **16**.

[4] Björnsson, H. and Venegas S.A. (1997): *A Manual of EOF and SVD Analysis of Climatic Data*. McGill University.

[5] Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Jenne, R., and Joseph, D. (1996): The NCEP/NCAR 40 Year Reanalysis Project. *Bulletin of the American Meteorological Society*.

[6] Košmelj, K. and Kastelec, D. (1996): *Statistične tabele*. Ljubljana: BF, Oddelek za agronomijo.

[7] Krzanowsky, W.J (1998): *An Introduction to Statistical Modelling*. London: Arnold.

[8] Lieth, H. (Ed.) (1974): *Phenology and Seasonality Modelling*. London: Chapman and Hall Lim.

[9] Maak, K. and von Storch, H. (1997): Statistical downscaling of monthly air temperature to the beginning of flowering of Galanthus nivalis L. in Northern Germany. *International Journal of Biometeorology*, **41**, 5-12

[10] Rummukainen, M. (1997): Methods for statistical downscaling of GCM simulations. *Reports Meteorology and Climatology*. Norrkoping, SMHI: 33.

[11] Schubert, S. (1998): Downscaling local extreme temperature changes in South-Eastern Australia from the CSIRO MARK2 GCM. *International Journal of Climatology*, **18**, 1419-1438.

[12] Simon, J.E., Chadwick, A.F., and Craker, L.E. (1984): Herbs: *An indexed Bibliography. 1971-1980. The Scientific Lliterature on Selected Herbs, and Aromatic and Medical Plants of the Temperature Zone*. Archon books.

[13] Spano, D., Cesaraccio, C., Duce, P., and Snyder, R.L. (1999): Phenological stages of natural species and their use as climate indicators. *International Journal of Biometeorology*, **42**, 124-133.

[14] von Storch, H., Zorita, E., and Cubash, U. (1993): Downscaling of global climate change estimates to regional scales: An aplication to Iberian rainfall in wintertime. *Journal of Climate*, **6**, 1161-1171.

[15] von Storch, H. (1995): Inconsistencies at the interface of climate impact studies and global climate research. *Meteorologische Zeitschrift*, **4**, 72-80.

[16] von Storch, H. and Zwiers, F.W. (1999): *Statistical Analysis in Climate Research*. Cambridge: Cambridge University Press.

[17] Wackernagel, H. (1998): Principal component analysis for autocorrelated data: A geostatistical perspective. Centre de Geostatistique - Ecole des Mines de Paris. *Technical Report*, **22G**.

[18] Walkovszky, A. (1998): Changes in phenology of the locust tree (Robinia pseudoacacia L.) in Hungary. *International Journal of Biometeorology*, **41**, 155-160.

[19] Wilby, R.L., Wigley, T.M.L. (1997): Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography*, **21**, 530-548.

[20] Zorita, E., von Storch, H. (1999): The analog method as simple statistical downscaling technique: comparison with more complicated methods. *Journal of Climate*, **12**, 2474-2489.