State - Space Discrimination and Clustering of Atmospheric Time Series Data Based on Kullback Information Measures

Thomas Bengtsson † and Joseph E. Cavanaugh ‡
† Statistics and Data Mining Department, Bell Labs, Murray Hill, NJ,
‡ Department of Biostatistics, University of Iowa, Iowa City, IA.

July 20, 2007

Abstract

Statistical problems in atmospheric science are frequently characterized by large spatio-temporal data sets and pose difficult challenges in classification and pattern recognition. Here, we consider the problem of identifying geographically homogeneous regions based on similarities in the temporal dynamics of weather patterns. Two disparity measures are proposed and applied to cluster time series of observed monthly temperatures from locations across Colorado, USA. The two disparity measures are based on state-space models, where the monthly temperature anomaly dynamics and seasonal variation are represented by latent processes. Our disparity measures produce clusters consistent with known atmospheric flow structures. In particular, the temporal anomaly pattern is related to the topography of Colorado, where, separated by the Continental Divide, the flow structures in the western and eastern parts of the state have different dynamics. The results further suggest that seasonal variation may be affected by locally-changing solar radiation levels primarily associated with elevation variations across the Rocky Mountains. The general methodology is outlined and developed in the Appendix. We conclude with a discussion of extensions to time varying and non-stationary systems.

Keywords: classification, pattern recognition, geostatistics, principal component analysis, principal oscillation pattern, state-space process.

1 Introduction

The goal of this work is to identify geographically homogeneous regions based on similarities in the temporal dynamics of weather patterns. To this end, we formulate two disparity measures for clustering time series of observed monthly temperatures, one based on anomaly dynamics and one based on seasonal component variation. Data is obtained through the National Climatic Data Center, which collects and archives US weather data for research and industrial users. We focus on mean monthly temperatures from weather stations located in Colorado, a state which comprises many different climatic zones along varying elevation contours, potentially related to topographical temperature clusters and patterns.

Cluster analysis is a commonly used tool for grouping weather events into climatologically homogeneous geographical regions (e.g., Gong and Richman 1995; Richman and Lamb 1985; Fovell and Fovell 1993), and for grouping time periods into clusters of homogeneous weather patterns (e.g., Alsop 1989). The application of classification techniques to atmospheric time series data is generally based on a decomposition of covariance structures using empirical orthogonal functions. Examples are given by Stone (1989) and Mo and Ghil (1988); see also Jolliffe (2002). However, this research typically does not directly account for the temporal associations of a series. In contrast, within the framework of the state-space model, we explicitly represent temporal associations using multivariate autoregressive models, and propose methods for allocation and separation based expressly on modeling temporal dynamics. More generally, our methodology allows clustering and discrimination according to unobserved structural components of a state-space process.

A related body of research exists in the statistical sciences, where discrimination and clustering of time series are based on assessing "distance" between a collection of observations. As in the general study of time series, procedures are developed in both the frequency and time domain; Shumway and Stoffer (2006) detail both approaches. Since the discrete Fourier transform asymptotically yields normally distributed and uncorrelated random variables, resulting in disparity functions that are straightforward to evaluate and investigate, most work in the statistical literature has focused on the frequency domain. This research includes developments by Liggett (1971), Shumway and Unger (1974), Dargahi-Noubary and Laycock (1981), Alagon (1989), Chaudhuri (1992), and Kakizawa, Shumway, and Taniguchi (1998). When dealing with multivariate

time series, a potential practical drawback of frequency domain methods is the sample variability of non-parametric spectral estimates. For time series of small to moderate length, the effects of such variability in the multivariate case will likely be pronounced, and may result in high miss-classification rates. An alternative approach is to assume (and possibly estimate) a lower-dimensional statistical model and evaluate the disparity measures in the time domain. Assuming univariate autoregressive moving-average (ARMA) models, Chan, Chinipardaz, and Cox (1996) derive a linear disparity function and its asymptotic distribution. Similarly relying on ARMA models, Gersch (1981) uses Kullback information measures to develop a classification procedure for stationary and locally stationary Gaussian time series. Related work include the methods of Melard and Roy (1983) and Coates and Diggle (1986).

For long series, the computational requirements of time domain approaches can be prohibitive as they typically require specification and inversion of high dimensional, possibly ill-conditioned matrices. The methods and analyses presented here rely on an existing or estimated state-space model, and address the heavy computational issues associated with evaluating disparity measures in the time domain through the Kalman filter recursions. Additionally, our approach allows for classification according to unobserved processes, a feature not available in maximum-likelihood based methodologies unless the data is pre-processed.

In the next section, we describe our problem, introduce our disparity measures, and use these tools to cluster monthly temperature records from across Colorado. We evaluate the propriety of the results using an alternative clustering approach derived from a principal component analysis. In the Appendix, we outline and develop our general methodology, which allows clustering and discrimination based on the latent structural components of a state-space process. Section 3 discusses extensions to systems with time dependent dynamics and to non-linear processes.

2 Methodology and Results

With the objective of identifying geographically homogeneous regions based on similarities in the temporal dynamics, two disparity measures are formulated and applied to cluster observed monthly temperatures from locations across Colorado, USA. Starting from the late-1800s, the complete data set contains daily measurements on a set of weather related variables from approximately 23.000 stations across the nation.

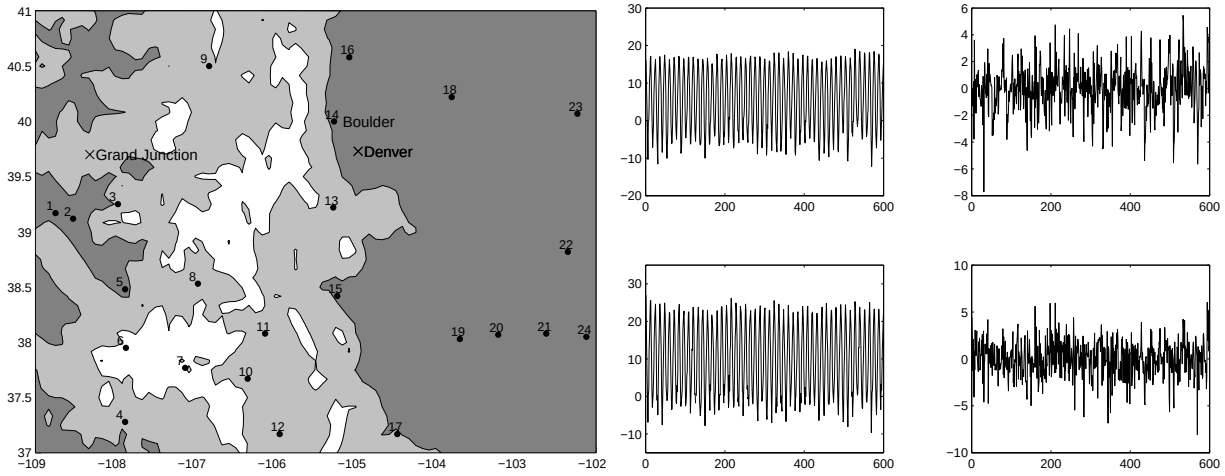

Figure 1: Left panel: Locations (lon/lat) of the 24 weather stations used in study. The shading represents elevation contours. Right panel: Covering the period 1947-1997, the series on the left show mean monthly temperatures (Celsius) for stations 22 (bottom) and 10 (top). Based on the same data, the series on the right depict deviations from monthly means for stations 22 and 10.

The complete records contain many missing observations, with numerous stations operating only for a brief time. Our focus is on mean monthly temperatures from stations located across Colorado, a state which comprises many different climatic zones along varying elevation contours - potentially defining topographical temperature clusters and patterns. The data is obtained through the National Climatic Data Center. For a detailed description of the data, including collection procedures and quality checks, we refer the reader to the webpage `http://dss.ucar.edu/datasets/ds510.0` (Scientific Computing Division-National Center for Atmospheric Research 2006).

Within Colorado we choose a subset of 24 (out of approximately 280) weather stations, selecting only those with at most 10% missing observations for the period 1895-1997. The left panel of Figure 1 shows the geographic locations of the weather stations used in the study. For reference, the cities of Denver, Boulder, and Grand Junction are located in the figure. Also depicted are three elevation contours: dark shade indicates elevations above 1000 meters (m), light shade indicates elevations above 2000m, and no shade indicates elevations above 3000m.

To illustrate the data, the mean monthly temperature of the last 50 years for stations 22 and 10 are depicted in the right panel of Figure 1 (bottom and top left). As expected, the monthly mean temperature series are dominated by a strong yearly cycle, with highest temperatures during summer months. However,

close inspection of Figure 1 shows slowly varying departures (anomalies) from the seasonal cycle. To give examples, the monthly temperature anomalies for stations 22 and 10, estimated by subtracting the individual monthly sample means from the data, are also shown in the right panel (bottom and top right). As can be seen, the estimated anomalies contain slowly varying modes. Subsection 2.3 provides nearest-neighbor clustering results targeting the anomaly and seasonal patterns for the mean monthly temperature series.

To further describe the data, Table 1 provides sample statistics for the western- and eastern-most stations (i.e., for stations 1-12 and 13-24, respectively). The overall sample means for the 24 stations varied between $.7^oC$ and 11.8^oC, with slightly warmer and more variable average temperatures on the plains. The distributions of mean monthly temperatures are approximately Gaussian, although some stations have slightly left skewed winter temperatures. With data from station 22, the histograms of Figure 2 provide examples of the distribution of monthly means.

Station#	1	2	3	4	5	6	7	8	9	10	11	12	mean
$\hat{\mu}_i$	9.5	11.5	8.1	7.8	9.7	3.5	0.7	2.3	3.7	5.6	6.4	5.3	6.2
$\hat{\sigma}_{\mathbf{Y}_i}$	9.6	9.7	9.0	8.2	9.0	6.9	9.0	10.0	9.5	7.9	8.4	9.0	8.9
N_{miss}	11	11	95	36	71	123	11	2	35	11	59	11	N/A
Station#	13	14	15	16	17	18	19	20	21	22	23	24	mean
$\hat{\mu}_i$	6.4	9.5	10.7	8.4	10.3	9.4	10.9	11.8	11.5	9.9	9.7	11.8	10.0
$\hat{\sigma}_{\mathbf{Y}_i}$	7.4	8.0	7.9	8.7	8.0	9.8	9.3	9.6	9.9	9.4	9.8	9.7	9.0
N_{miss}	11	59	11	11	11	16	11	11	13	71	59	11	N/A

Table 1: Sample means, $\hat{\mu}_i$, and standard deviations, $\hat{\sigma}_{\mathbf{Y}_i}$, for stations $i = 1, \ldots, 24$. N_{miss} denotes the number of missing observations out of a total of 1236 months for the period 1895-1997.

2.1 State-space data model

We fit the monthly data using an additive, structural state-space model. For station i, the model represents the monthly mean as a sum of an overall constant mean, a seasonal component, a monthly temperature anomaly, and a white noise term. The structural components are denoted respectively by μ_i, s_{it}, a_{it}, and ϵ_{it}, where s_{it} and a_{it} are latent processes. Thus, with y_{it} representing the observed mean temperature for station i and month t, we have $y_{it} = \mu_i + s_{it} + a_{it} + \epsilon_{it}$.

Differing mean temperature levels among the various stations could easily be used to delineate different geographic regions; however, as our main interest is in the dynamics of the monthly series, we remove the overall mean of each series. For simplicity, this is done approximately by subtracting sample means,

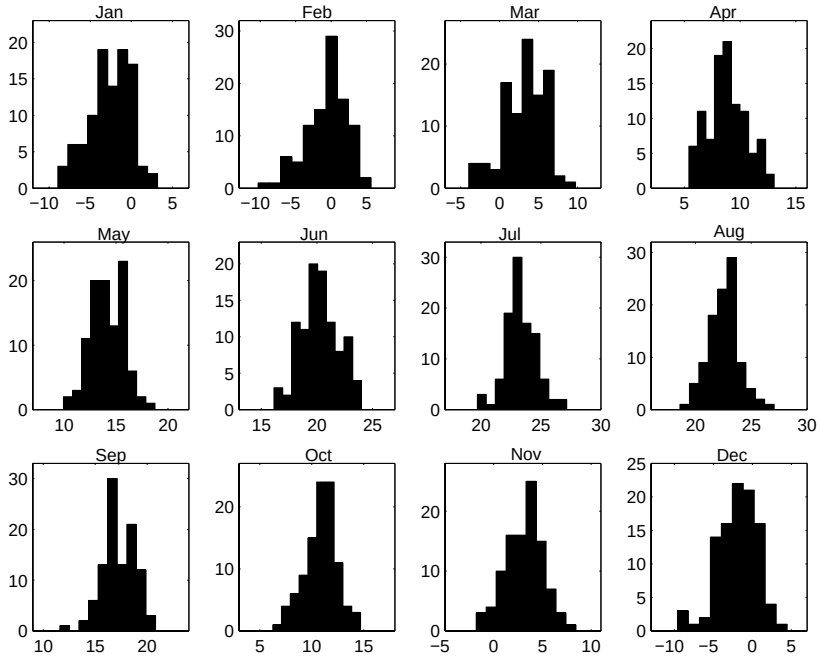

Figure 2: Histograms of the monthly mean temperatures (^{o}C) for station 22.

yielding the data model $\tilde{y}_{it} = y_{it} - \hat{\mu}_i = s_{it} + a_{it} + \epsilon_{it}$, where $\hat{\mu}_i$ is the overall sample mean of station i (see Table 1). The seasonal components are thought of as slowly varying processes, and are modeled by letting $(\sum_{j=t-11}^{t} s_{ij}) = \delta_{it} \overset{iid}{\sim} N(0, \sigma^2_{\delta_i})$ (see, e.g., Janacek and Swift 1993), and the monthly temperature anomalies are modeled by AR(1) processes, $a_{it} = \phi_i a_{i,t-1} + \eta_{it}$, with $\eta_{it} \overset{iid}{\sim} N(0, \sigma^2_{\eta_i})$. Here, following our previous assessment of the distribution of monthly means (see Figure 2), the state-error processes are taken as Gaussian. The white noise term is viewed as contributing variability that is unexplained by the structural components. We let $\epsilon_{it} \overset{iid}{\sim} N(0, \sigma^2_{\epsilon_i})$.

For this data, geographically nearby stations covary, implying correlated state error processes; i.e., for two stations $i \neq j$ and $cov(y_{it}, y_{jt}) \neq 0$, we have $cov(\delta_{it}, \delta_{jt}) \neq 0$ and (or) $cov(\eta_{it}, \eta_{jt}) \neq 0$. However, since the spatial domain of our study includes varying elevation and climate zones, no prior knowledge of the spatial structures of the noise processes is assumed. (In fact, the purpose of our study is to identify spatial and temporal patterns.) Hence, for modeling purposes, we set $cov(\eta_{it}, \eta_{jt}) = 0$ and $cov(\delta_{it}, \delta_{jt}) = 0, i \neq j, \forall t$. Ignoring non-zero across-station error covariances will not affect the clustering results, but will yield inefficient (yet unbiased) parameter estimation and state prediction. Finally, to "separate" seasonal and anomaly

patterns, we assume $cov(\delta_{it}, \eta_{jt}) = 0, \forall i, j, t$.

For computational convenience, we represent the preceding monthly temperature model in general state-space form. Our notation and setup for the linear state-space model is similar to that of Shumway and Stoffer (2006, p 324-325). The general form of the model allows for straightforward implementation of the Kalman filter and smoother recursions, which recover the latent processes, as well as the expectation-maximization (EM) algorithm, which provides maximum likelihood (ML) estimates of the model parameters (see Shumway and Stoffer 1982; and Shumway and Stoffer 2006, p 342-344).

A linear state-space process $\mathbf{y}_t$ is represented by two sets of equations:

$$\mathbf{y}_t = \mathbf{A}_t \mathbf{x}_t + \mathbf{v}_t, \tag{1}$$

$$\mathbf{x}_t = \mathbf{\Phi} \mathbf{x}_{t-1} + \mathbf{w}_t, \quad t = 1, \ldots, N. \tag{2}$$

In the observation equation (1), the design matrix $\mathbf{A}_t$ relates the unobserved state vector $\mathbf{x}_t$ to the observed vector $\mathbf{y}_t$. In the state equation (2), the transition matrix $\mathbf{\Phi}$ relates $\mathbf{x}_t$ to its previous value $\mathbf{x}_{t-1}$ via an autoregression. The vectors $\mathbf{v}_t$ and $\mathbf{w}_t$ represent zero-mean white noise processes with covariance structures $\mathbf{R}$ and $\mathbf{Q}$, respectively. The model assumes a prior distribution for $\mathbf{x}_0$ with $E(\mathbf{x}_0) = \boldsymbol{\mu}$ and $cov(\mathbf{x}_0) = \mathbf{V}$, and $\mathbf{x}_0$ is taken to be uncorrelated with $\mathbf{v}_t$ and $\mathbf{w}_t$ for all t. Normality is often assumed for both error processes as well as for $\mathbf{x}_0$. We will let $\mathbf{\Theta} = \{\boldsymbol{\mu}, \mathbf{V}, \mathbf{\Phi}, \mathbf{Q}, \mathbf{R}\}$ denote the set of parameters for the model defined in (1) and (2), and let $\mathbf{Y} = [\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_N']'$ and $\mathbf{X} = [\mathbf{x}_0', \mathbf{x}_1', \ldots, \mathbf{x}_N']'$ represent the vectors of observed data and unobserved states.

For station i, let $\mathbf{x}_{it} = [a_{it}, s_{it}, \ldots, s_{i,t-10}]'$. With this state vector, the structural model of monthly mean temperature for station i is put into state-space form as follows. For the observation equation (1), we have

$$\tilde{y}_{it} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} a_{it} \\ s_{it} \\ \vdots \\ s_{i,t-10} \end{pmatrix} + \epsilon_{it}, \tag{3}$$

and for the state equation (2),

$$\begin{pmatrix} a_{it} \\ s_{it} \\ \vdots \\ s_{i,t-10} \end{pmatrix} = \begin{pmatrix} \phi_i & 0 & \cdots & \cdots & 0 \\ 0 & -1 & \cdots & \cdots & -1 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} a_{i,t-1} \\ s_{i,t-1} \\ \vdots \\ s_{i,t-11} \end{pmatrix} + \begin{pmatrix} \eta_{it} \\ \delta_{it} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{4}$$

Here, $\mathbf{A}_{it}$ and $\mathbf{\Phi}_i$ are as defined in (3) and (4), and $\mathbf{\Theta}_i = \{\phi_i, \sigma^2_{\eta_i}, \sigma^2_{\delta_i}, \sigma^2_{\epsilon_i}\}$.

2.2 Parameter estimates and model fit

The four rows of Tables 2 and 3 show parameter estimates for the 12 western- and 12 eastern-most stations in the study. A clear pattern among the parameters can be seen by comparing the entries of the tables. For example, stations at higher elevation, i.e., the western-most stations, have smaller autoregressive coefficients, with a mean of .620 (.867 for the eastern stations), and higher anomaly innovations variances, with a mean of 1.19 (.139). Clearly, the temporal dynamics of monthly temperature are associated with spatial location.

ID	1	2	3	4	5	6	7	8	9	10	11	12	mean
$\hat{\phi}$	.562	.442	.650	.730	.645	.771	.677	.582	.765	.463	.668	.492	.620
$\hat{\sigma}^2_\eta$	1.60	2.16	.735	.470	.867	.354	1.12	2.13	.434	1.56	1.23	1.69	1.19
$\hat{\sigma}^2_\delta \times 10^2$	.150	.029	.368	.001	.322	.641	.273	.340	.077	.058	.207	.221	.200
$\hat{\sigma}^2_\epsilon$	1.43	1.12	1.71	1.58	1.59	1.58	1.71	1.31	2.16	1.10	1.38	1.00	1.47

Table 2: Parameter estimates for the 12 *western*-most stations.

To verify that the model in (3) and (4) fits the data adequately, residual analyses were performed for each of the 24 temperature series. Satisfactory fits were obtained for all station data. As examples, Figure 3 depicts autocorrelation and QQ-plots for the residuals of stations 10 and 22. As indicated in the autocorrelation plots, the seasonal component has effectively been removed and there does not exist strong evidence against a white-noise (null) hypothesis on the residuals. Although the QQ-plots indicate slightly left-skewed residuals, severe departures from normality are not evident.

ID	13	14	15	16	17	18	19	20	21	22	23	24	mean
$\hat{\phi}$	.858	.830	.830	.998	.831	.902	.900	.939	.705	.862	.998	.745	.867
$\hat{\sigma}^2_\eta$	.226	.152	.112	.002	.114	.166	.055	.087	.418	.124	.001	.216	.139
$\hat{\sigma}^2_\delta \times 10^2$	.001	.093	.064	.003	.135	.125	.086	.002	.212	.083	.303	.115	.100
$\hat{\sigma}^2_\epsilon$	2.43	3.61	2.97	3.83	2.59	4.01	3.17	3.46	3.26	3.62	4.28	3.40	3.39

Table 3: Parameter estimates for the 12 *eastern*-most stations.

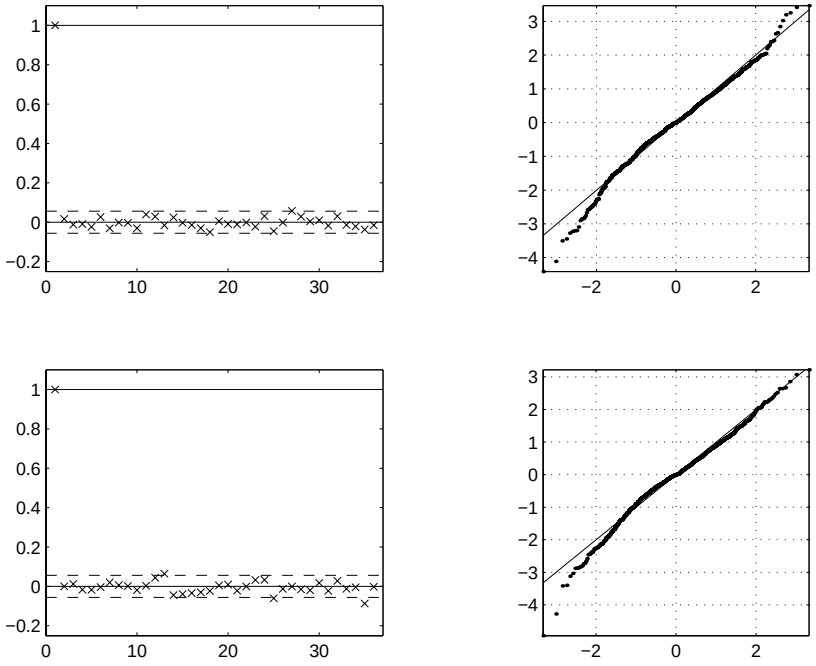

Figure 3: Autocorrelation plots and QQ-plots of residuals for stations 22 (top) and 10 (bottom). The plots are based on sample sizes of 1225 and 1165 data points, respectively. Approximate 95% upper and lower points (dashed) for the null distribution are shown in the autocorrelation plots.

2.3 Discrepancy measures and clustering results

Using the ML-parameter estimates of Tables 2 and 3, which were obtained using the EM algorithm, two discrepancy measures are formulated, calculated, and used for clustering the temperature data. The first discrepancy measure matches the densities of the anomaly process, and the second discrepancy matches the densities of the seasonal cycle. To provide a context for our results, a principal component analysis based on estimated monthly anomalies is performed.

To introduce our discrepancy measures we define the following notation. Let N_i denote the sample size for station i, and let $\mathbf{Y}_i = [\tilde{y}_{i1}, \tilde{y}_{i2}, \ldots, \tilde{y}_{i,N_i}]'$ denote the (mean adjusted) time series for this station. Also, let $\mathbf{X}_i = [\mathbf{x}'_{i1}, \mathbf{x}'_{i2}, \ldots, \mathbf{x}'_{i,N_i}]'$ and $\mathbf{a}_i = [a_{i1}, a_{i2}, \ldots, a_{i,N_i}]'$ represent the latent state and anomaly processes. Let $\mathbf{S}_{it}(0)$, $\mathbf{S}_{it}(1)$, and $\mathbf{S}_{i,t-1}(0)$ represent the smoothing matrices with respect to $\mathbf{Y}_i$ and $\mathbf{\Theta}_i$, as defined by (11) and (12) in Subsection A.2 of the Appendix, and let $S^{11}_{it}(0)$, $S^{11}_{it}(1)$, and $S^{11}_{i,t-1}(0)$ denote the upper left-hand entries of these matrices.

The first discrepancy measure is based solely on the autoregressive anomaly process as defined in (4). For

each station i, under the assumption of independent state-error processes, $f(a_{it}, s_{it}|\mathbf{\Theta}_i) = f(a_{it}|\mathbf{\Theta}_i)f(s_{it}|\mathbf{\Theta}_i)$. Then, following the developments in Subsection A.3 of the Appendix, we define a discrepancy measure based on the anomaly part of the state vector by considering the Kullback information (Kullback 1968) for the anomaly densities:

$$d^a(\mathbf{Y}_i, \mathbf{\Theta}_i; \mathbf{\Theta}_j) = \int \log \frac{f(\mathbf{a}_i|\mathbf{\Theta}_i)}{f(\mathbf{a}_i|\mathbf{\Theta}_j)} f(\mathbf{X}_i|\mathbf{Y}_i, \mathbf{\Theta}_i) d\mathbf{X}_i. \tag{5}$$

A computational formula for $d^a(\mathbf{Y}_i, \mathbf{\Theta}_i; \mathbf{\Theta}_j)$ is derived by considering the two terms in (15). Following the developments in the Appendix that lead to (14), we obtain

$$\begin{aligned} d^a(\mathbf{Y}_i, \mathbf{\Theta}_i; \mathbf{\Theta}_j) = &- \frac{N_i}{2}(\log \sigma^2_{\eta_i} - \log \sigma^2_{\eta_j}) \\ &-\{\frac{1}{2\sigma^2_{\eta_i}}[S^{11}_{it}(0) - 2\phi_i S^{11}_{it}(1) + \phi_i^2 S^{11}_{i,t-1}(0)] - \frac{1}{2\sigma^2_{\eta_j}}[S^{11}_{it}(0) - 2\phi_j S^{11}_{it}(1) + \phi_j^2 S^{11}_{i,t-1}(0)]\}. \end{aligned}$$

Clustering algorithms require the use of symmetrized discrepancy measures (see, for instance, Kakizawa, Shumway, and Taniguchi 1998). Thus, based on (5), we define a form of the J-divergence (Kullback 1968) that accounts for the different lengths of each series in the data sets by averaging over time:

$$\bar{J}^a(\mathbf{Y}_i, \mathbf{\Theta}_i; \mathbf{Y}_j, \mathbf{\Theta}_j) = N_i^{-1} d^a(\mathbf{Y}_i, \mathbf{\Theta}_i; \mathbf{\Theta}_j) + N_j^{-1} d^a(\mathbf{Y}_j, \mathbf{\Theta}_j; \mathbf{\Theta}_i).$$

Employing output from the EM-algorithm, including the ML parameter estimates, the sample $\bar{J}^a$-divergence reduces to

$$\begin{aligned} \bar{J}^a(\mathbf{Y}_i, \hat{\mathbf{\Theta}}_i; \mathbf{Y}_j, \hat{\mathbf{\Theta}}_j) = &\frac{1}{2N_j\hat{\sigma}^2_{\eta_i}}[S^{11}_{jt}(0) - 2\hat{\phi}_i S^{11}_{jt}(1) + \hat{\phi}_i^2 S^{11}_{j,t-1}(0)] \\ &+ \frac{1}{2N_i\hat{\sigma}^2_{\eta_j}}[S^{11}_{it}(0) - 2\hat{\phi}_j S^{11}_{it}(1) + \hat{\phi}_j^2 S^{11}_{i,t-1}(0)] - 1. \end{aligned} \tag{6}$$

Using the parameter estimates of Tables 2 and 3 in calculating $\bar{J}^a(\mathbf{Y}_i, \hat{\mathbf{\Theta}}_i; \mathbf{Y}_j, \hat{\mathbf{\Theta}}_j)$, $i, j = 1, \ldots, 24$, produced the nearest neighbor results of Figure 4. (Connecting nearest neighbors produces results similar to those obtained by a minimum spanning tree, which recursively connects the closest centroids.) It can be seen from the plot, with the exception of station 21, that all stations have nearest neighbors within their respective east-west geographic region. (We remind the reader that light colors correspond to higher elevations, and that the Rocky Mountains, which pass through the western region of the state, are represented by the light grey and white shadings.) As seen in the plot, the two cluster structures are distinctly separated along the

mid-longitude elevation contours. Results from three different hierarchical cluster methods, Ward's, single linkage, and complete linkage, confirm the depicted visual clusters.

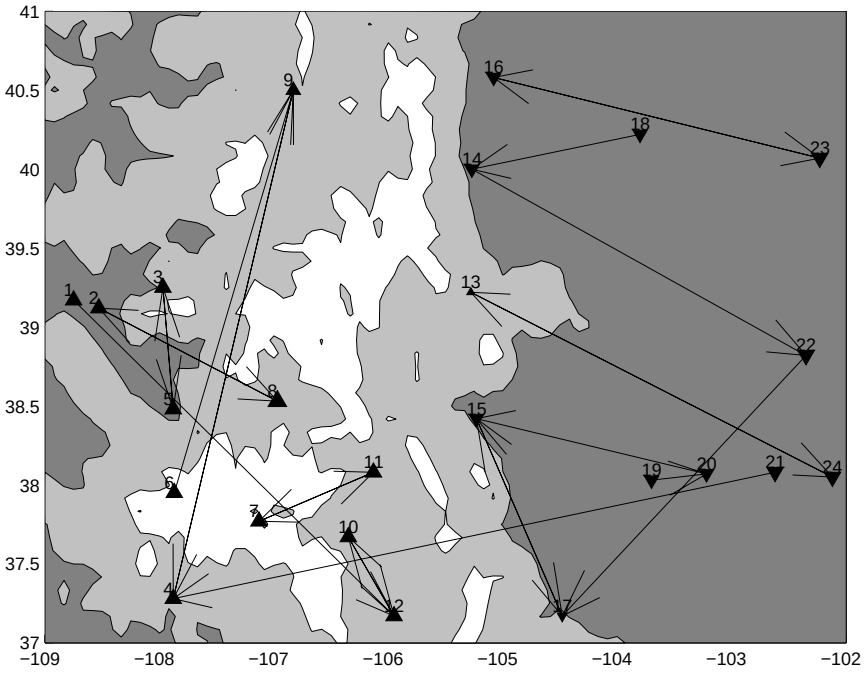

Figure 4: Nearest neighbors based on $\bar{J}^a(\mathbf{Y}_i, \hat{\mathbf{\Theta}}_i; \mathbf{Y}_j, \hat{\mathbf{\Theta}}_j)$. Neighbors are indicated by arrows: e.g., station 9 is the nearest neighbor of station 6. Based on the covariance matrix of the anomalies, negative factor loadings from the second PC are shown by downward pointing triangles (positive-upwards), with loading magnitude proportional to triangle size.

To provide a comparison with existing methodology we performed a principal component analysis (PCA) targeting the anomaly process. It should be noted that a number of fairly involved approaches to empirical orthogonal function analysis of spatio-temporal data sets (e.g., principal oscillation pattern analysis) have been developed in recent years, each method addressing slightly different aspects of the covariance structures of atmospheric data (see, e.g., Jolliffe 2002, and Wikle 2002). However, as these methods are rather complex, we compare our results only with those produced by decomposition of the correlation matrix of the scalar observations.

To obtain factor loadings appropriate for comparison with the clusters depicted in Figure 4, the PCA is based on the estimated anomaly series. As with the anomaly data depicted in Figure 1, the anomaly series are calculated by subtraction of station-specific monthly means from the original data. The correlation matrix was chosen as a basis for the principal components because anomaly variability differed somewhat from

ID	1	2	3	4	5	6	7	8	9	10	11	12	mean
$\hat{e}_{ia}$	.226	.229	.221	.183	.222	.162	.208	.242	.155	.203	.206	.209	.206
ID	13	14	15	16	17	18	19	20	21	22	23	24	mean
$\hat{e}_{ia}$	.004	-.180	-.170	-.184	-.097	-.180	-.206	-.216	-.265	-.242	-.256	-.255	-.187

Table 4: Sample loadings, $\hat{e}_{ia}$, based on the second principal component. The loadings are based on the correlation matrix of the anomaly data.

station to station. (For this data, the factor loadings based on the covariance matrix differ only marginally from those based on the correlation matrix.) To address numerous missing values, element-wise correlation estimates were used to define the sample correlation matrix. The factor loadings from the first principal component (not depicted), accounting for 67.7% of total variability, are spatially unstructured (i.e., positive and of approximately equal size) and indicative of a climatologically homogeneous, or limited, geographical region.

The factor loadings from the second principal component, accounting for approximately 10% of total variability, are given in Table 4. With means of .206 and -.187, the factor loadings for the western and eastern stations are clearly delineated by longitude. The spatial distribution of the factor loadings is also depicted in Figure 4, with a negative factor loading indicated by a downward pointing triangle (upward for a positive loading), and with loading magnitude proportional to triangle size. For this data, the anomaly pattern of the second principal component is identical to the cluster pattern derived from $\bar{J}^a(\mathbf{Y}_i, \hat{\mathbf{\Theta}}_i; \mathbf{Y}_j, \hat{\mathbf{\Theta}}_j)$.

As previously mentioned, the temporal (and spatial) temperature anomaly pattern is related to the topography of Colorado. Separated by the Continental Divide, the flow patterns in the western and eastern parts of the state have different dynamics, here reflected by different parameter estimates, during strong winter and summer anomalies. However, the spatial pattern among the factor loadings may at least partially be due to the fact that a limited area is studied, potentially producing boundary effects determined by large scale (global) anomaly temperature patterns (Richman 1986). Nevertheless, the results of both methods confirm the existence of clearly defined anomaly patterns. In fact, as no direct information about the temporal covariance structures is obtained by the principal component analysis performed here, the two methods target different information yet produce very similar results.

The second discrepancy measure is based on the seasonal process defined in (4). By developments similar to those producing the computational formula (6), targeting the anomaly process, one can show that the

sample discrepancy for the seasonal process is given by

$$\bar{J}^s(\mathbf{Y}_i, \hat{\mathbf{\Theta}}_i; \mathbf{Y}_j, \hat{\mathbf{\Theta}}_j) = \frac{1}{2}\left(\frac{\hat{\sigma}^2_{\delta_i}}{\hat{\sigma}^2_{\delta_j}} + \frac{\hat{\sigma}^2_{\delta_j}}{\hat{\sigma}^2_{\delta_i}}\right) - 1. \tag{7}$$

It should be noted that the simple form of equation (7) is partly due to the fact that the EM-algorithm is used to produce output. We note further the dependence of $\bar{J}^s(\mathbf{Y}_i, \hat{\mathbf{\Theta}}_i; \mathbf{Y}_j, \hat{\mathbf{\Theta}}_j)$ on the estimated state noise variances $\hat{\sigma}^2_{\delta_i}$ and $\hat{\sigma}^2_{\delta_j}$, the parameters defining the covariance structures of the seasonal processes.

Nearest-neighbor clustering results for the seasonal process, based on $\bar{J}^s(\mathbf{Y}_i, \hat{\mathbf{\Theta}}_i; \mathbf{Y}_j, \hat{\mathbf{\Theta}}_j)$, are shown in Figure 5. Although somewhat complex, east-west and north-south patterns are discernable. To more clearly identify potential clusters, the discrepancy matrix produced by evaluation of $\bar{J}^s(\mathbf{Y}_i, \hat{\mathbf{\Theta}}_i; \mathbf{Y}_j, \hat{\mathbf{\Theta}}_j), i, j = 1, \ldots, 24$, was subjected to Ward's, single linkage, and complete linkage clustering procedures (e.g., Hartigan 1975; Gordon 1999). As the results from the three methods are similar, each essentially identifying four clusters, we discuss only the results of Ward's method. Detailed results from Ward's method are provided by the dendrogram in Figure 6. As seen in the plot, the identified clusters are given by stations 3, 13; stations 5, 7; stations 4, 6, 8, 10, 11, 16, 19, 20, 21, 22; and stations 1, 2, 9, 12, 14, 15, 17, 18, 23, 24.

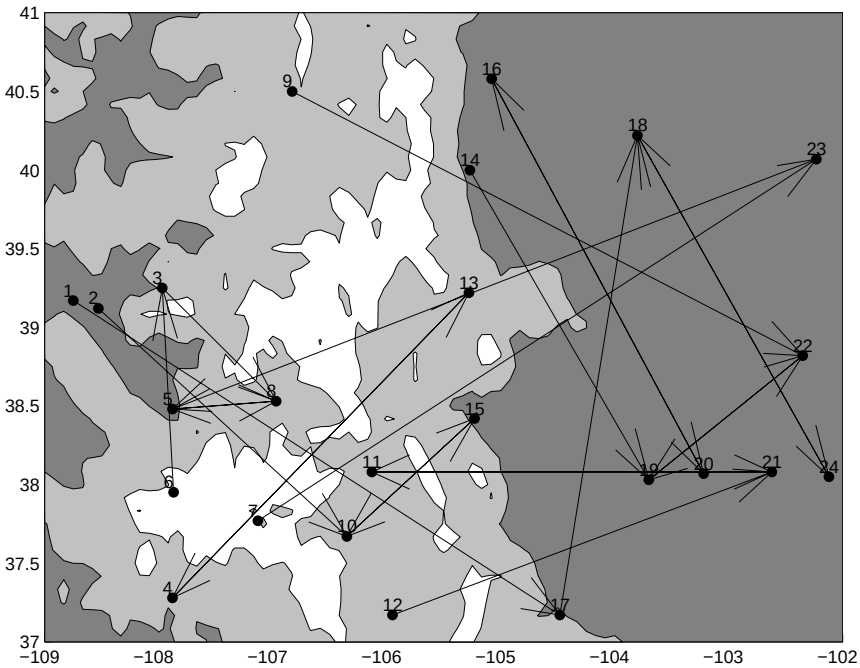

Figure 5: Nearest neighbors based on $\bar{J}^s(\mathbf{Y}_i, \hat{\mathbf{\Theta}}_i; \mathbf{Y}_j, \hat{\mathbf{\Theta}}_j)$.

As the seasonal phase is primarily attributable to latitude, only the south-north grouping of cluster three

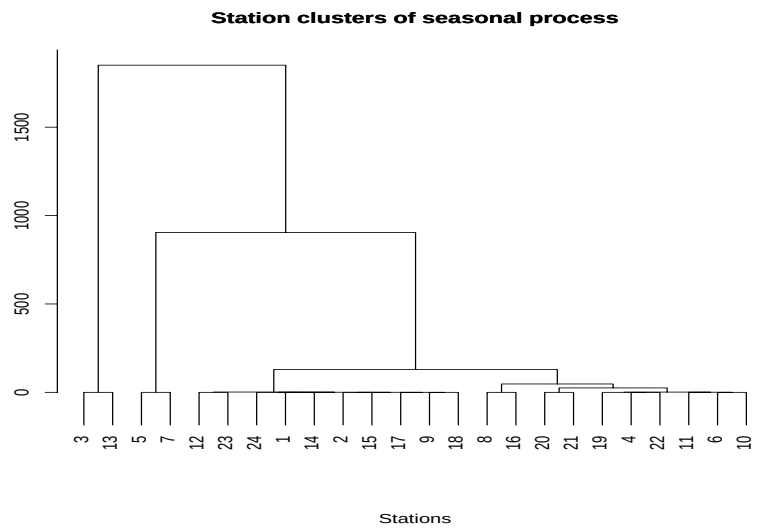

Figure 6: Dendrogram based on Ward's method.

can be clearly associated to obvious spatial variability in the seasonal cycle. The east-west grouping of cluster four may potentially result from different albedo (solar radiation reflection) levels between the mountains and the plains. Also, the topography of the Rocky Mountains may limit amplitude changes in the seasonal cycle. Absent physical explanation, the clusters of stations 3 and 13, and stations 5 and 7 remain obscure, but the existence of such clusters will likely diminish when a larger spatial domain, with more variability in the seasonal cycle, can be observed. In fact, the domain under consideration here is likely too small to produce robust seasonal clusters.

There is no simple method with which to compare the obtained seasonal clusters. One possibility is to fit a spatially varying mean to each of the 12 months of the year, and treat the residual seasonal deviations as data. However, this procedure will likely be dominated by regions of similar mean structure and yield little insight about the dynamics of the seasonal cycle. Another possibility is to perform a PCA that directly takes into account the seasonal cycle, e.g. periodically extended empirical orthogonal function analysis, as in Kim and Wu (1999). As mentioned, these methods are somewhat involved and will not be pursued here.

Next we discuss generalizations of the disparity function and suggest extensions to time varying and non-linear systems.

3 Extensions to Time Varying Systems

The measures developed in Subsection A.2 and A.3 of the Appendix can be extended to time varying systems. One useful example of a time varying state-space process allows the state transition matrix to change: i.e., in (2) we may consider $\mathbf{x}_t = \mathbf{\Phi}_t \mathbf{x}_{t-1} + \mathbf{w}_t$. For example, the time series data considered in Section 2 could be modeled by letting the autoregressive component ϕ vary with month. Thus, depending on the month, let ϕ_t be from $\{\phi_{jan}, \phi_{feb}, \ldots, \phi_{dec}\}$. Another important class of time varying processes is obtained by letting the system depend non-linearly on the state, e.g., consider $\mathbf{x}_t = \mathbf{\Phi}(\mathbf{x}_{t-1}) + \mathbf{w}_t$, where $\mathbf{\Phi}(\mathbf{x}_{t-1})$ is a function of $\mathbf{x}_{t-1}$. The extended Kalman filter (*cf.* Jazwinski 1970) can then be used to approximately evaluate the statistics of the Kalman recursions, with transition matrix at time t given by $\mathbf{\Phi}_t = \partial \mathbf{\Phi}(c)/\partial c$ evaluated at the forecast $c = \tilde{\mathbf{x}}_t(\mathbf{\Theta}, \mathbf{Y}^{t-1})$. To apply the discriminant and clustering methods presented here to time varying systems, expressions based on $\mathbf{\Phi}$ must be adapted to depend on the parameter set $\{\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_N\}$ (see the computational formulae (13) and (14) in Subsections A.2 and A.3). The extension yields somewhat more complicated expressions, but the development is mostly a matter of bookkeeping.

It should be noted that strongly non-linear processes with rapid error growth, i.e., chaotic systems or any system which depends sensitively on initial conditions, cannot be modeled using linear approximations (Miller, Ghil, and Gauthiez 1994). In such scenarios, Monte-Carlo based filtering methods are often used to produce (approximate) samples from the posterior distribution. Assuming such samples $\tilde{\mathbf{X}}_\star^k \sim f(\mathbf{X}|\mathbf{Y}_\star, \mathbf{\Theta}_\star),\ k = 1, 2, \ldots, m$, are available, the general discrepancy measure (8) in Subsection A.2 could (in principle) by estimated by $\widehat{d}(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}, \mathbf{\Theta}) = m^{-1} \sum_{i=1}^m \log[f(\tilde{\mathbf{X}}_\star^k|\mathbf{Y}_\star, \mathbf{\Theta}_\star)/f(\tilde{\mathbf{X}}_\star^k|\mathbf{Y}, \mathbf{\Theta})]$. The ensemble Kalman filter (and smoother), a widely used data assimilation method in the atmospheric community (e.g., Houtekamer and Mitchell 1998; Bengtsson, Nychka, and Snyder 2003), emulates the extended Kalman filter by propagating the first two moments of the filter distribution, and can be adapted to evaluate the Monte Carlo based discrepancy even for strongly non-linear systems.

4 Summary

Using Kullback information, new disparity measures for classification of state-space processes are used to cluster monthly temperature records from stations across Colorado, USA. The results confirm known at-

mospheric flow patterns, and agree with those produced by a principal component analysis. Further, the results suggest that the seasonal cycle may be locally affected by albedo levels associated with the elevation variations of Colorado. The disparity measures are model-based and provide a general, structural approach to pattern recognition in time series analysis. In particular, classification according to unobserved latent components, hitherto not directly possible using existing methods, is feasible. Efficient computational evaluation of the disparity measures is provided by output from the Kalman smoother. The work is concluded with a discussion of extensions of our methods to time varying systems. A detailed development of the general methodology is presented in the Appendix.

ACKNOWLEDGEMENTS: The authors would like to thank Dr. J. Tribbia and Dr. D. Nychka, both NCAR, for valuable insights and suggestions. They would also like to thank the deputy editor and a referee for useful suggestions which helped to strengthen the organization and exposition.

A Appendix: Development of Disparity Measures for State-Space Processes

In this Appendix, we present the relevant technical results that pertain to our clustering procedures. Subsection A.1 introduces the linear state-space framework, and describes notation for the required prediction quantities. We then present new disparity measures for gauging similarity between time series data modeled within this framework. Specifically, Subsection A.2 introduces a general state-space divergence measure for time series data based on the Kullback information, and Subsection A.3 adapts this measure to focus more specifically on the latent state process. Formulae that permit exact computational evaluation of the proposed measures are also derived. Finally, in Subsection A.4, we discuss the use of these formulae in discrimination and clustering algorithms.

A.1 Linear state-space model

The state-space model is a general dynamic linear model and subsumes a rich class of specialized models for time series and stochastic processes (Jazwinski 1970; Brockwell and Davis 1991). A primary tool in engineering and signal processing (Anderson and Moore 1979; Chen and Liu 2000), the state-space model is also widely applied in other disciplines, e.g., economics (Harvey 1989; Durbin and Koopman 2001), medicine (Jones 1993), and the atmospheric sciences (Wikle and Cressie 1999; Houtekamer and Mitchell 2001).

Let $\mathbf{Y}^t = [\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_t']'$ and $\mathbf{X}^t = [\mathbf{x}_0', \mathbf{x}_1', \ldots, \mathbf{x}_t']'$ represent the vectors of observed data and unobserved states through time t ($t = 1, \ldots, N$). Of primary concern in state-space modeling is prediction and recovery of the unobserved states in $\mathbf{X}^N$. Under the model assumptions listed in Subsection 2.1, prediction of $\mathbf{x}_t$ using the conditional mean $E(\mathbf{x}_t|\boldsymbol{\Theta}, \mathbf{Y}^k)$, here denoted $\tilde{\mathbf{x}}_t(\boldsymbol{\Theta}, \mathbf{Y}^k)$, yields the best predictor of $\mathbf{x}_t$ given $\mathbf{Y}^k$, $1 \leq k, t \leq N$. For $k \leq t$, the predictors are obtained recursively using the Kalman filter (Kalman 1960): the one-step ahead predictors $\tilde{\mathbf{x}}_t(\boldsymbol{\Theta}, \mathbf{Y}^{t-1})$ and the filters $\tilde{\mathbf{x}}_t(\boldsymbol{\Theta}, \mathbf{Y}^t)$ are obtained through the forward Kalman filter recursions, which also generate the innovations $e_t(\boldsymbol{\Theta}, \mathbf{Y}^t) = \mathbf{y}_t - \mathbf{A}_t E(\mathbf{x}_t|\boldsymbol{\Theta}, \mathbf{Y}^{t-1})$ along with the covariance matrices $\boldsymbol{\Sigma}_t(\boldsymbol{\Theta}) = E\{e_t(\boldsymbol{\Theta}, \mathbf{Y}^t)e_t(\boldsymbol{\Theta}, \mathbf{Y}^t)'|\boldsymbol{\Theta}, \mathbf{Y}^{t-1}\}$. For $t \leq k \equiv N$, the smoothers $\tilde{\mathbf{x}}_t(\boldsymbol{\Theta}, \mathbf{Y}^N)$ are obtained recursively using the backwards smoothing algorithms (Anderson and Moore 1979), which also produce the error covariances $\mathbf{P}_t^N(\boldsymbol{\Theta}) = E\{[\mathbf{x}_t - \tilde{\mathbf{x}}_t(\boldsymbol{\Theta}, \mathbf{Y}^N)][\mathbf{x}_t - \tilde{\mathbf{x}}_t(\boldsymbol{\Theta}, \mathbf{Y}^N)]'|\boldsymbol{\Theta}, \mathbf{Y}^N\}$.

Estimates of the unknown parameters in $\boldsymbol{\Theta}$ are usually obtained using maximum likelihood. We use the EM-algorithm method of ML estimation, adapted to the state-space framework by Shumway and Stoffer (1982), as our disparity functions can be obtained as by-products of the output.

In what follows, for notational simplicity, we drop the superscripts on the data and state vectors and write $\mathbf{Y}$ and $\mathbf{X}$ to mean $\mathbf{Y}^N$ and $\mathbf{X}^N$, respectively. (This simplified notational convention was also followed in the body of the paper.) Additionally, we assume that N is the same for all of the data and state vectors introduced in the subsequent development.

A.2 General state-space divergence measure

As the primary interest of state-space modeling is often the unobserved process $\mathbf{X}$, a natural way to compare state-space processes is to match the mechanisms generating the states. Following Cavanaugh and Johnson (1999), we use an information theoretic approach to develop a disparity measure that gauges the similarity of two state-space processes by comparing posterior (conditional) densities of the unobserved states. (Based on these results, a measure targeting the marginal densities of the states is developed in the next subsection.)

We wish to assess the disparity between two state-space processes $\mathbf{Y}_\star$ and $\mathbf{Y}$, each process generated according to (1) and (2) with parameter structures $\boldsymbol{\Theta}_\star$ and $\boldsymbol{\Theta}$, respectively. Our general discrepancy function assesses the similarity of the unobserved states of $\mathbf{Y}_\star$ and $\mathbf{Y}$ by comparing $f(\mathbf{X}|\mathbf{Y}_\star, \boldsymbol{\Theta}_\star)$ with $f(\mathbf{X}|\mathbf{Y}, \boldsymbol{\Theta})$. Since $\mathbf{X}$ is unobserved, we consider the comparison suggested by the Kullback information (Kullback 1968), also known as the I-divergence:

$$d(\mathbf{Y}_\star, \boldsymbol{\Theta}_\star; \mathbf{Y}, \boldsymbol{\Theta}) = \int \log \frac{f(\mathbf{X}|\mathbf{Y}_\star, \boldsymbol{\Theta}_\star)}{f(\mathbf{X}|\mathbf{Y}, \boldsymbol{\Theta})} f(\mathbf{X}|\mathbf{Y}_\star, \boldsymbol{\Theta}_\star) d\mathbf{X}. \tag{8}$$

In addition to contrasting parameters (the objective of the traditional Kullback information based on marginal as opposed to conditional densities), the preceding divergence also compares data. To provide further insight about $d(\mathbf{Y}_\star, \boldsymbol{\Theta}_\star; \mathbf{Y}, \boldsymbol{\Theta})$, the next proposition delineates the defined divergence in terms of the smoothed representations of $\mathbf{Y}_\star$ and $\mathbf{Y}$, and the posterior covariance matrices $cov(\mathbf{X}|\boldsymbol{\Theta}_\star, \mathbf{Y}_\star)$ and $cov(\mathbf{X}|\boldsymbol{\Theta}, \mathbf{Y})$.

Proposition 1. Let $\tilde{\mathbf{x}}_\star = E(\mathbf{X}|\boldsymbol{\Theta}_\star, \mathbf{Y}_\star)$ and $\tilde{\mathbf{x}} = E(\mathbf{X}|\boldsymbol{\Theta}, \mathbf{Y})$ contain the complete sets of smoothed values of $\mathbf{Y}_\star$ and $\mathbf{Y}$ under the models defined by $\boldsymbol{\Theta}_\star$ and $\boldsymbol{\Theta}$. Also, let $\boldsymbol{\Sigma}_\star = cov(\mathbf{X}|\boldsymbol{\Theta}_\star, \mathbf{Y}_\star)$ and $\boldsymbol{\Sigma} = cov(\mathbf{X}|\boldsymbol{\Theta}, \mathbf{Y})$.

Then, with p = dimension of $\mathbf{x_t}$,

$$d(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}, \mathbf{\Theta}) = \frac{1}{2}\{\log\left(\frac{|\mathbf{\Sigma}|}{|\mathbf{\Sigma}_\star|}\right) + tr(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\star) + (\tilde{\mathbf{x}}_\star - \tilde{\mathbf{x}})'\mathbf{\Sigma}^{-1}(\tilde{\mathbf{x}}_\star - \tilde{\mathbf{x}}) - Np\}. \tag{9}$$

Proposition 1 follows by noting that $d(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}, \mathbf{\Theta})$ is the I-divergence of two multivariate normal distributions (see Kullback 1968, p 306).

As indicated by (9), $d(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}, \mathbf{\Theta})$ compares posterior means as well as posterior covariance matrices. The point-wise comparison of posterior means provides evaluation of the measurement scales and potential phase shifts between $\mathbf{y}_{t,\star}$ and $\mathbf{y}_t$. In contrast, the posterior covariance structures are not dependent on $\mathbf{Y}_\star$ or $\mathbf{Y}$ under Gaussian model assumptions, and will reflect only differences between parameter structures. Thus, assuming $\mathbf{\Theta}_\star$ and $\mathbf{\Theta}$ represent the same measurement process, the temporal dynamics of $\mathbf{x}_t$ under the two models are indirectly contrasted by comparison of the posterior covariance matrices $\mathbf{\Sigma}_\star$ and $\mathbf{\Sigma}$.

Since the posterior covariance matrices $\mathbf{\Sigma}_\star$ and $\mathbf{\Sigma}$ are assumed high-dimensional (at least $N \times N$), and furthermore require specification and inversion, Proposition 1 is not computationally useful. An exact computational formula that allows for efficient evaluation of $d(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}, \mathbf{\Theta})$ is given in Proposition 2.

Proposition 2. Let $L(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y})$ represent the complete-data likelihood of $\mathbf{Y}$ and $\mathbf{X}$, and define $Q(\mathbf{\Theta}_\star, \mathbf{Y}_\star; \mathbf{\Theta}, \mathbf{Y}) = E\{\log L(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y})|\mathbf{\Theta}_\star, \mathbf{Y}_\star\}$. Then,

$$d(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}, \mathbf{\Theta}) = [Q(\mathbf{\Theta}_\star, \mathbf{Y}_\star|\mathbf{\Theta}_\star, \mathbf{Y}_\star) - Q(\mathbf{\Theta}, \mathbf{Y}|\mathbf{\Theta}_\star, \mathbf{Y}_\star)] + [\log L(\mathbf{\Theta}|\mathbf{Y}) - \log L(\mathbf{\Theta}_\star|\mathbf{Y}_\star)]. \tag{10}$$

Proof. The result follows by noting that $f(\mathbf{X}|\mathbf{Y}, \mathbf{\Theta}) = f(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta})/f(\mathbf{Y}|\mathbf{\Theta})$. The first two terms in (10) can be calculated using the Kalman smoother, while the latter two terms are obtained through calculation of likelihoods.

Using output from the backward smoothing recursions, Shumway and Stoffer (1982, p 256-7) and Shumway and Stoffer (2006, p 343) present an explicit form for $Q(\mathbf{\Theta}, \mathbf{Y}|\mathbf{\Theta}_\star, \mathbf{Y}_\star)$. With $\tilde{\mathbf{x}}_t(\mathbf{\Theta}, \mathbf{Y}) = E(\mathbf{x}_t|\mathbf{\Theta}, \mathbf{Y})$, and $\mathbf{P}^N_{t,t-1}(\mathbf{\Theta}) = E\{[\mathbf{x}_t - \tilde{\mathbf{x}}_t(\mathbf{\Theta}, \mathbf{Y})][\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_{t-1}(\mathbf{\Theta}, \mathbf{Y})]'|\mathbf{\Theta}, \mathbf{Y}\}$ representing the cross-covariance

$cov(\mathbf{x}_t, \mathbf{x}_{t-1}|\boldsymbol{\Theta}, \mathbf{Y})$, Shumway and Stoffer (1982) argue

$$
\begin{aligned}
2Q(\boldsymbol{\Theta}, \mathbf{Y}|\boldsymbol{\Theta}_\star, \mathbf{Y}_\star) = & \\
& - \log|\mathbf{V}| - tr(\mathbf{V}^{-1}\{\mathbf{P}_0^N(\boldsymbol{\Theta}_\star) + [\tilde{\mathbf{x}}_0(\boldsymbol{\Theta}, \mathbf{Y}_\star) - \boldsymbol{\mu}][\tilde{\mathbf{x}}_0(\boldsymbol{\Theta}, \mathbf{Y}_\star) - \boldsymbol{\mu}]'\}) \\
& - N\log|\mathbf{Q}| - tr\{\mathbf{Q}^{-1}[\mathbf{S}_t(0) - \mathbf{S}_t(1)\boldsymbol{\Phi}' - \boldsymbol{\Phi}\mathbf{S}_t(1)' + \boldsymbol{\Phi}\mathbf{S}_{t-1}(0)\boldsymbol{\Phi}']\} \\
& - N\log|\mathbf{R}| - tr(\mathbf{R}^{-1}\{\sum_{t=1}^{N}[\mathbf{y}_t - \mathbf{A}_t\tilde{\mathbf{x}}_t(\boldsymbol{\Theta}_\star, \mathbf{Y}_\star)][\mathbf{y}_t - \mathbf{A}_t\tilde{\mathbf{x}}_t(\boldsymbol{\Theta}_\star, \mathbf{Y}_\star)]' + \mathbf{A}_t\mathbf{P}_t^N(\boldsymbol{\Theta}_\star)\mathbf{A}_t'\}),
\end{aligned}
$$

where

$$
\mathbf{S}_t(j) = \sum_{t=1}^{N}[\mathbf{P}_{t,t-j}^N(\boldsymbol{\Theta}_\star) + \tilde{\mathbf{x}}_t(\boldsymbol{\Theta}_\star, \mathbf{Y}_\star)\tilde{\mathbf{x}}_{t-j}(\boldsymbol{\Theta}_\star, \mathbf{Y}_\star)'], \tag{11}
$$

and

$$
\mathbf{S}_{t-1}(0) = \sum_{t=1}^{N}[\mathbf{P}_{t-1}^N(\boldsymbol{\Theta}_\star) + \tilde{\mathbf{x}}_{t-1}(\boldsymbol{\Theta}_\star, \mathbf{Y}_\star)\tilde{\mathbf{x}}_{t-1}(\boldsymbol{\Theta}_\star, \mathbf{Y}_\star)']. \tag{12}
$$

Then, with $\tilde{\mathbf{x}}_t = E(\mathbf{x}_t|\boldsymbol{\Theta}_\star, \mathbf{Y}_\star)$, one can show

$$
\begin{aligned}
d(\mathbf{Y}_\star, \boldsymbol{\Theta}_\star; \mathbf{Y}, \boldsymbol{\Theta}) = & -\frac{N}{2}(\log|\mathbf{Q}_\star| - \log|\mathbf{Q}|) \\
& - \frac{1}{2}tr(\{\mathbf{Q}_\star^{-1}[\mathbf{S}_t(0) - \mathbf{S}_t(1)\boldsymbol{\Phi}_\star' - \boldsymbol{\Phi}_\star\mathbf{S}_t(1)' + \boldsymbol{\Phi}_\star\mathbf{S}_{t-1}(0)\boldsymbol{\Phi}_\star']\} \\
& \qquad - \{\mathbf{Q}^{-1}[\mathbf{S}_t(0) - \mathbf{S}_t(1)\boldsymbol{\Phi}' - \boldsymbol{\Phi}\mathbf{S}_t(1)' + \boldsymbol{\Phi}\mathbf{S}_{t-1}(0)\boldsymbol{\Phi}']\}) \\
& - \frac{N}{2}(\log|\mathbf{R}_\star| - \log|\mathbf{R}|) \\
& - \frac{1}{2}tr(\{\mathbf{R}_\star^{-1}\sum_{t=1}^{N}[(\mathbf{y}_{t,\star} - \mathbf{A}_t\tilde{\mathbf{x}}_t)(\mathbf{y}_{t,\star} - \mathbf{A}_t\tilde{\mathbf{x}}_t)' + \mathbf{A}_t\mathbf{P}_t^N(\boldsymbol{\Theta}_\star)\mathbf{A}_t']\} \\
& \qquad - \{\mathbf{R}^{-1}\sum_{t=1}^{N}[(\mathbf{y}_t - \mathbf{A}_t\tilde{\mathbf{x}}_t)(\mathbf{y}_t - \mathbf{A}_t\tilde{\mathbf{x}}_t)' + \mathbf{A}_t\mathbf{P}_t^N(\boldsymbol{\Theta}_\star)\mathbf{A}_t']\}) \\
& + [\log L(\boldsymbol{\Theta}|\mathbf{Y}) - \log L(\boldsymbol{\Theta}_\star|\mathbf{Y}_\star)],
\end{aligned} \tag{13}
$$

where the smoothing matrices $\mathbf{S}_t(0)$, $\mathbf{S}_t(1)$, and $\mathbf{S}_{t-1}(0)$ are evaluated using $\mathbf{Y}_\star$ and the model defined by $\boldsymbol{\Theta}_\star$. In the computational formula, the contribution from the prior density of $\mathbf{x}_0$ on $d(\mathbf{Y}_\star, \boldsymbol{\Theta}_\star; \mathbf{Y}, \boldsymbol{\Theta})$ is ignored.

Evaluation of $d(\mathbf{Y}_\star, \boldsymbol{\Theta}_\star; \mathbf{Y}, \boldsymbol{\Theta})$ simplifies somewhat when ML parameter estimates are used to calculate the involved smoothing quantities. For instance, with $p = \text{dimension}(\mathbf{x}_t)$, $q = \text{dimension}(\mathbf{y}_t)$, and with $\hat{\boldsymbol{\Theta}}_\star$ estimated from $\mathbf{Y}_\star$, using the EM algorithm: $Np \equiv tr\{\hat{\mathbf{Q}}_\star^{-1}[\mathbf{S}_t(0) - \mathbf{S}_t(1)\hat{\boldsymbol{\Phi}}_\star' - \hat{\boldsymbol{\Phi}}_\star\mathbf{S}_t(1)' + \hat{\boldsymbol{\Phi}}_\star\mathbf{S}_{t-1}(0)\hat{\boldsymbol{\Phi}}_\star']\}$, and $Nq \equiv tr[\hat{\mathbf{R}}_\star^{-1}\sum_{t=1}^{N}\{[\mathbf{y}_{t,\star} - \mathbf{A}_t\tilde{\mathbf{x}}_t][\mathbf{y}_{t,\star} - \mathbf{A}_t\tilde{\mathbf{x}}_t]' + \mathbf{A}_t\mathbf{P}_t^N(\hat{\boldsymbol{\Theta}}_\star)\mathbf{A}_t'\}]$. •

As can be seen from the computational formula in Proposition 2, the defined I-divergence results in discrimination directly dependent on the observation equation, a potentially undesirable effect if the primary interest is the state process. However, as shown next, the divergence in (8) can be used to obtain an estimate of the Kullback information between $f(\mathbf{X}|\boldsymbol{\Theta}_\star)$ and $f(\mathbf{X}|\boldsymbol{\Theta})$.

A.3 State disparity measure

In many cases it may be desirable to have a disparity function based solely on the state process, even for state-space series with different measurement processes. Consider then, as an alternative to $d(\mathbf{Y}_\star, \boldsymbol{\Theta}_\star; \mathbf{Y}, \boldsymbol{\Theta})$, defining a discrepancy that targets only the state densities $f(\mathbf{X}|\boldsymbol{\Theta}_\star)$ and $f(\mathbf{X}|\boldsymbol{\Theta})$:

$$d^{state}(\mathbf{Y}_\star, \boldsymbol{\Theta}_\star; \boldsymbol{\Theta}) = \int \log \frac{f(\mathbf{X}|\boldsymbol{\Theta}_\star)}{f(\mathbf{X}|\boldsymbol{\Theta})} f(\mathbf{X}|\mathbf{Y}_\star, \boldsymbol{\Theta}_\star) d\mathbf{X}. \tag{14}$$

As stated by the next result, $d^{state}(\mathbf{Y}_\star, \boldsymbol{\Theta}_\star; \boldsymbol{\Theta})$ provides an unbiased estimate of the Kullback information between $f(\mathbf{X}|\boldsymbol{\Theta}_\star)$ and $f(\mathbf{X}|\boldsymbol{\Theta})$.

Proposition 3.

$$\int d^{state}(\mathbf{Y}_\star, \boldsymbol{\Theta}_\star; \boldsymbol{\Theta}) f(\mathbf{Y}_\star|\boldsymbol{\Theta}_\star) d\mathbf{Y}_\star = \int \log \frac{f(\mathbf{X}|\boldsymbol{\Theta}_\star)}{f(\mathbf{X}|\boldsymbol{\Theta})} f(\mathbf{X}|\boldsymbol{\Theta}_\star) d\mathbf{X}.$$

Proof. Since the support of $\mathbf{X}$ and $\mathbf{Y}_\star$ are independent, we have

$$\begin{aligned}
\int d^{state}(\mathbf{Y}_\star, \boldsymbol{\Theta}_\star; \boldsymbol{\Theta}) f(\mathbf{Y}_\star|\boldsymbol{\Theta}_\star) d\mathbf{Y}_\star &= \int_{\mathbf{Y}_\star} \left\{ \int_{\mathbf{X}} \log \Big[\frac{f(\mathbf{X}|\boldsymbol{\Theta}_\star)}{f(\mathbf{X}|\boldsymbol{\Theta})} \Big] \frac{f(\mathbf{Y}_\star|\mathbf{X}, \boldsymbol{\Theta}_\star) f(\mathbf{X}|\boldsymbol{\Theta}_\star)}{f(\mathbf{Y}_\star|\boldsymbol{\Theta}_\star)} d\mathbf{X} \right\} f(\mathbf{Y}_\star|\boldsymbol{\Theta}_\star) d\mathbf{Y}_\star \\
&= \int_{\mathbf{X}} \log \Big[\frac{f(\mathbf{X}|\boldsymbol{\Theta}_\star)}{f(\mathbf{X}|\boldsymbol{\Theta})} \Big] f(\mathbf{X}|\boldsymbol{\Theta}_\star) \left\{ \int_{\mathbf{Y}_\star} f(\mathbf{Y}_\star|\mathbf{X}, \boldsymbol{\Theta}_\star) d\mathbf{Y}_\star \right\} d\mathbf{X} \\
&= \int_{\mathbf{X}} \log \Big[\frac{f(\mathbf{X}|\boldsymbol{\Theta}_\star)}{f(\mathbf{X}|\boldsymbol{\Theta})} \Big] f(\mathbf{X}|\boldsymbol{\Theta}_\star) d\mathbf{X}.
\end{aligned}$$

The result also follows readily from the law of iterated expectations. •

Based on reasoning similar to that used in establishing Proposition 2, the computational formula for $d^{state}(\mathbf{Y}_\star, \boldsymbol{\Theta}_\star; \boldsymbol{\Theta})$ is given by

$$\begin{aligned}
d^{state}(\mathbf{Y}_\star, \boldsymbol{\Theta}_\star; \boldsymbol{\Theta}) = &- \frac{N}{2}(\log|\mathbf{Q}_\star| - \log|\mathbf{Q}|) \\
&- \frac{1}{2} tr(\{\mathbf{Q}_\star^{-1}[\mathbf{S}_t(0) - \mathbf{S}_t(1)\boldsymbol{\Phi}_\star' - \boldsymbol{\Phi}_\star \mathbf{S}_t(1)' + \boldsymbol{\Phi}_\star \mathbf{S}_{t-1}(0)\boldsymbol{\Phi}_\star']\} \\
&\quad - tr\{\mathbf{Q}^{-1}[\mathbf{S}_t(0) - \mathbf{S}_t(1)\boldsymbol{\Phi}' - \boldsymbol{\Phi}\mathbf{S}_t(1)' + \boldsymbol{\Phi}\mathbf{S}_{t-1}(0)\boldsymbol{\Phi}']\}),
\end{aligned} \tag{15}$$

where the smoothing matrices $\mathbf{S}_t(0), \mathbf{S}_t(1)$, and $\mathbf{S}_{t-1}(0)$, are as previously defined.

Proposition 3 does not guarantee the positivity of $d^{state}(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{\Theta})$, but when the EM algorithm is used to estimate $\mathbf{\Theta}_\star$ from $\mathbf{Y}_\star$ and to smooth the data, it can be verified that $d^{state}(\mathbf{Y}_\star, \hat{\mathbf{\Theta}}_\star; \mathbf{\Theta}) \geq 0$. Thus, the sample state disparity measure can be used as a pseudo-distance measure for discrimination and clustering purposes.

Using an argument similar to that of Proposition 3, one can develop a divergence that targets a subset of the state vector. That is, suppose the state vector can be partitioned, e.g., $\mathbf{X} = [\mathbf{X}_1', \mathbf{X}_2']'$, and is such that $f(\mathbf{X}|\mathbf{\Theta}_\star) = f(\mathbf{X}_1|\mathbf{\Theta}_\star)f(\mathbf{X}_2|\mathbf{\Theta}_\star)$, then the Kullback information divergences

$$\int \log \frac{f(\mathbf{X}_1|\mathbf{\Theta}_\star)}{f(\mathbf{X}_1|\mathbf{\Theta})} f(\mathbf{X}_1|\mathbf{\Theta}_\star) d\mathbf{X}_1 \qquad \text{and} \qquad \int \log \frac{f(\mathbf{X}_2|\mathbf{\Theta}_\star)}{f(\mathbf{X}_2|\mathbf{\Theta})} f(\mathbf{X}_2|\mathbf{\Theta}_\star) d\mathbf{X}_2$$

can be targeted using output from $d^{state}(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{\Theta})$. The use of such discrepancy functions is illustrated in the application.

Importantly, Proposition 3 and the computational formula for $d^{state}(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{\Theta})$ provide a basis for data-dependent classification targeting only the density of the unobserved processes, a result that cannot be obtained by a likelihood-based method in the state-space setting.

A.4 Discrimination and clustering using $d(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}, \mathbf{\Theta})$ and $d^{state}(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{\Theta})$

A flexible non-parametric discriminant rule with asymptotically optimal properties is given by the nearest neighbor rule (Cover and Hart 1967). The nearest neighbor rule assigns to an unclassified sample point the classification of the nearest, with respect to some distance/similarity measure, of a set of previously labeled sample points. Because of lack of distributional assumptions and simple use of $d(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}, \mathbf{\Theta})$ and $d^{state}(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{\Theta})$, we utilize the nearest neighbor rule to classify a new time series into a previously labeled population (Gersch 1981).

Suppose we wish to determine population membership of $\mathbf{Y}_\star$ (generated under $\mathbf{\Theta}_\star$), and have available a set of n time series, $\mathbf{S}(n) = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$, each series drawn from one of k mutually exclusive populations, uniquely defined by the parameters $\mathbf{\Omega}(k) = \{\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_k\}$. Then, given $(\mathbf{Y}_\star, \mathbf{\Theta}_\star)$, the nearest neighbor of $\mathbf{Y}_\star$ is a function of both $\mathbf{Y}$ and $\mathbf{\Theta}$, and evaluation of $d(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}, \mathbf{\Theta})$ requires selection of $\mathbf{Y}$ from $\mathbf{S}(n)$ and $\mathbf{\Theta}$ from $\mathbf{\Omega}(k)$. If $\mathbf{Y}_j \in \mathbf{S}(n)$ is generated under the state-space model defined by $\mathbf{\Theta}_i \in \mathbf{\Omega}(k)$, the

obvious choice is to select the data-parameter pair $(\mathbf{Y}, \mathbf{\Theta}) = (\mathbf{Y}_j, \mathbf{\Theta}_i)$. However, this choice results in simultaneous data-parameter comparison, and is sensitive to scale and phase differences between $\mathbf{Y}_\star$ and $\mathbf{Y}_j$. Thus, for discrimination based on population characteristics (here synonymous with parameters), we propose the discrepancy resulting from comparing the smoothed density of $\mathbf{Y}_\star$ under the models defined by $\mathbf{\Theta}_\star$ and $\mathbf{\Theta}_j$. With $(\mathbf{Y}, \mathbf{\Theta}) = (\mathbf{Y}_\star, \mathbf{\Theta}_j)$ in (8), we define the discrepancy $d_1(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}_\star, \mathbf{\Theta}_j)$, with nearest neighbor allocation of $\mathbf{Y}_\star$ based on $\min_{\{1 \le j \le k\}} d_1(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}_\star, \mathbf{\Theta}_j)$. Similarly, with $\mathbf{\Theta} = \mathbf{\Theta}_j$ in (14), we define $d^{state}(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{\Theta}_j)$, with allocation of $\mathbf{Y}_\star$ according to $\min_{\{1 \le j \le k\}} d^{state}(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{\Theta}_j)$. The symmetric J-divergences defined by

$$J_1(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}_\star, \mathbf{\Theta}_j) = d_1(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}_\star, \mathbf{\Theta}_j) + d_1(\mathbf{Y}_\star, \mathbf{\Theta}_j; \mathbf{Y}_\star, \mathbf{\Theta}_\star), \text{ or} \tag{16}$$

$$J^{state}(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}_\star, \mathbf{\Theta}_j) = d^{state}(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{\Theta}_j) + d^{state}(\mathbf{Y}_\star, \mathbf{\Theta}_j; \mathbf{\Theta}_\star), \tag{17}$$

can also be used for classification of $\mathbf{Y}_\star$.

For clustering purposes we assume $n = k$. Then, the symmetric disparity measures

$$J_1(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}_j, \mathbf{\Theta}_j) = d_1(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}_\star, \mathbf{\Theta}_j) + d_1(\mathbf{Y}_j, \mathbf{\Theta}_j; \mathbf{Y}_j, \mathbf{\Theta}_\star), \text{ or} \tag{18}$$

$$J^{state}(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}_j, \mathbf{\Theta}_j) = d^{state}(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{\Theta}_j) + d^{state}(\mathbf{Y}_j, \mathbf{\Theta}_j; \mathbf{\Theta}_\star), \tag{19}$$

may be used to define a $k \times k$ distance matrix, $\mathbf{D}_{S(k)}$, reflecting the similarity of the series contained in $\mathbf{S}(k)$. Clustering of the series in $\mathbf{S}(k)$ into homogeneous subgroups may then be performed by decomposition of $\mathbf{D}_{S(k)}$ using classical partitioning procedures (e.g., Johnson and Wichern 1992). (Note the difference between (16) and (18) and between (17) and (19).)

In our application, we use two variants of $J^{state}(\mathbf{Y}_\star, \mathbf{\Theta}_\star; \mathbf{Y}_j, \mathbf{\Theta}_j)$ for clustering our monthly temperature time series: one based on the anomaly process and one based on the seasonal process.

References

Alagon, J. (1989). Spectral discrimination for two groups of time series. *Journal of Time Series Analysis 10*, 203–214.

Alsop, T. (1989). The natural seasons of western Oregon and Washington. *Journal of Climate 2*, 888–896.

Anderson, D. and J. Moore (1979). *Optimal Filtering*. Englewood Cliffs.

Bengtsson, T., D. Nychka, and C. Snyder (2003). Toward a nonlinear ensemble filter for high-dimensional systems. *Journal of Geophysical Research 108*(D24), 8775.

Brockwell, P. and R. Davis (1991). *Time Series: Theory and Methods* (2 ed.). Springer-Verlag.

Cavanaugh, J. and W. Johnson (1999). Assessing the predictive influence of cases in a state-space process. *Biometrika 86*, 183–190.

Chan, H., R. Chinipardaz, and T. Cox (1996). Discrimination of AR, MA and ARMA time series models. *Communications in Statistics: Theory and Methods 25*, 1247–1260.

Chaudhuri, G. (1992). Linear discriminant function for complex normal time series. *Statistics and Probability Letters 15*, 277–279.

Chen, R. and J. Liu (2000). Mixture Kalman filters. *Journal of the Royal Statistical Society B62*, 493–508.

Coates, D. and P. Diggle (1986). Tests for comparing two estimated spectral densities. *Journal of Time Series Analysis 7*, 7–20.

Cover, T. and P. Hart (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory 13*, 21–27.

Dargahi-Noubary, G. and P. Laycock (1981). Spectral ratio discriminants and information theory. *Journal of Time Series Analysis 2*, 71–86.

Durbin, J. and S. Koopman (2001). *Time Series Analysis by State-Space Methods*. Oxford University Press.

Fovell, R. and M. Fovell (1993). Climate zones of the conterminous United States defined using cluster analysis. *Journal of Climate 6*, 2103–2135.

Gersch, W. (1981). *In Applied Time Series Analysis II*, Nearest neighbor rule classification of stationary and nonstationary time series, pp. 221–270. Academic Press.

Gong, X. and M. Richman (1995). On the application of cluster analysis to growing season precipitation data in North America east of the Rockies. *Journal of Climate 8*, 897–931.

Gordon, A. D. (1999). *Classification*. London: Chapman and Hall / CRC.

Hartigan, J. A. (1975). *Clustering Algorithms*. New York: Wiley.

Harvey, A. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Houtekamer, P. and H. Mitchell (1998). Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review 126*, 796–811.

Houtekamer, P. and H. Mitchell (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review 129*, 123–137.

Janacek, G. and L. Swift (1993). *Time Series: Forecasting, Simulation, and Applications*. Ellis Horwood.

Jazwinski, A. (1970). *Stochastic Processes and Filtering Theory*. Academic Press.

Johnson, R. and D. Wichern (1992). *Applied Multivariate Statistical Analysis* (3 ed.). Prentice-Hall.

Jolliffe, I. (2002). *Principal Component Analysis* (2 ed.). Springer-Verlag.

Jones, R. (1993). *Longitudinal Data with Serial Correlation: A State-Space Approach*. Chapman & Hall.

Kakizawa, Y., R. Shumway, and M. Taniguchi (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association 93*, 328–340.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions ASME Journal of Basic Engineering 9*, 35–45.

Kim, K.-Y. and Q. Wu (1999). A comparison study of EOF techniques: Analysis of nonstationary data with periodic statistics. *Journal of Climate 12*, 185–199.

Kullback, S. (1968). *Information Theory and Statistics*. Dover.

Liggett, W. (1971). On the asymptotic optimality of spectral analysis for testing hypotheses about time series. *Annals of Mathematical Statistics 42*, 1348–58.

Melard, G. and R. Roy (1983). Testing for homogeneity and stability of time series. In *Proceedings of the American Statistical Association: Business and Economic Statistics Section*.

Miller, R., M. Ghil, and F. Gauthiez (1994). Advanced data assimilation in strongly nonlinear dynamical systems. *Journal of Atmospheric Sciences 51*, 1037–1056.

Mo, K. and M. Ghil (1988). Cluster analysis of multiple planetary flow regimes. *Journal of Geophysical Research 93*(D9), 10927–10952.

Richman, M. (1986). Rotation of principal components. *Journal of Climatology 6*, 293–335.

Richman, M. and J. Lamb (1985). Climatic pattern analysis of 3- and 7-day summer rainfall in the central United States: Some methodological considerations and a regionalization. *Journal of Climate and Applied Meterology 24*, 1325–1343.

Scientific Computing Division-National Center for Atmospheric Research (2006). NCDC TD3200 U.S. Coopertaive summary of day, 1890(1948)-cont.

Shumway, R. and D. Stoffer (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis 3*, 253–264.

Shumway, R. and D. Stoffer (2006). *Time Series Analysis and Its Applications* (2 ed.). Springer-Verlag.

Shumway, R. and A. Unger (1974). Linear discriminant functions for stationary time series. *Journal of the American Statistical Association 69*, 948–956.

Stone, R. C. (1989). Weather types at Brisbane, Queensland: An example of the use of principal components and cluster analysis. *International Journal of Climatology 9*, 3–32.

Wikle, C. (2002). *Encyclopedia of Life Support Systems*, Spatio-temporal models in climatology. EOLSS Publishers Co.

Wikle, C. and N. Cressie (1999). A dimension reduced approach to space-time Kalman filtering. *Biometrika 86*, 815–829.