International Journal of Climatology

Downscaling Middle East Rainfall using a Support Vector Machine and Hidden Markov Model

Journal:	*International Journal of Climatology*
Manuscript ID:	draft
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Samuels, Rana; Tel Aviv University, Geophysics Robertson, Andrew; Columbia University, IRI/Earth Institute Khalil, Abedalrazq; Columbia University, Earth and Environmental Engineering Lall, Upmanu; Columbia University, Earth and Environmental Engineering
Keywords:	hidden markov model, downscaling, precipitation, Middle East

Downscaling Middle East Rainfall using a Support Vector Machine and Hidden Markov Model

Rana Samuels[1,2], Andrew W. Robertson[3], Abedalrazq Khalil [1], Upmanu Lall[1,3]

(1) Columbia University, Department of Earth and Environmental Engineering, USA

(2) Tel Aviv University, Department of Geophysics and Planetary Sciences, Israel

(3) International Research Institute for Climate and Society (IRI), Columbia University, New York, USA

Corresponding Author:

Rana Samuels

ranas@post.tau.ac.il

tel: +972-3-640-9120

fax: +972-3-640-9282

ABSTRACT

A support vector machine (SVM) is combined with a non-homogeneous hidden Markov model (NHMM) to downscale daily station rainfall sequences over Israel from global atmospheric reanalysis data. The selected atmospheric fields can be extracted from general circulation models, making the SVM-NHMM approach potentially useful for downscaling climate predictions. For the October–March wet season, seasonally averaged rainfall variability was captured with an anomaly correlation skill of 0.89 over the period 1980 to 2000, using a model trained on a previous period. The frequency of 7-day dry spells and the length of the longest dry spell per season were reproduced correlations of 0.82 and 0.71, respectively. The SVM-NHMM approach is shown to outperform the PCA-NHMM approach used in previous studies, in which principal component analysis (PCA) is used for linear dimension reduction of the atmospheric predictors in place of the SVM. The NHMM hidden states generated are seen to correspond to synoptic systems prevalent in the region including the Cyprus low, a low pressure system centered over Cyprus and Turkey.

1. INTRODUCTION

Effectively simulating daily precipitation variability is important for different scales of water resource management and planning including crop choice and investment, aquifer management and water infrastructure investment. For rain-fed agriculture, the economic viability of the crops is critically affected by the farmer's ability to cope with within-season and longer-term climate variability..

It is increasingly realized that both global and local information is necessary to adequately simulate and predict rainfall variability at a given location over various time scales. Over the past 10 years, General Circulations Models (GCMs) and Regional Circulation Models (RCMs) have become increasingly effective and efficient at seasonal prediction of sea surface temperatures (SSTs), wind patterns, and near-surface air temperature and precipitation (Goddard et al., 2001; Sun et al., 2005). However, as the spatial resolution of GCMs is in the hundreds of kilometers, the predicted rainfall is not directly comparable to rain gauge data that are typically used by hydrologic and crop models; RCMs embedded over a subdomain of a GCM use GCM output as boundary conditions. They can be expensive and cumbersome to run as part of a multi-model ensemble prediction system and may still lead to significant biases in local daily rainfall statistics, that arise due to either flawed boundary conditions passed on by the GCM, or to flaws in the parameterization of physical hydrologic processes. In short, even for the GCM-RCM approach for climate model downscaling, the final rainfall output requires statistical processing to be useful.

Statistical models such as Markov processes and neural networks using only local data for rainfall simulation have been in use for some time. Weather generators using Markov chains have been used for decades to simulate stochastic local daily rainfall sequences based on the statistical characteristics of observed rainfall at that location (Gabriel and Neumann, 1962; Mason, 2004; Richardson, 1981). By integrating information from the GCMs and RCMs with appropriate local statistical models it is, in principle, possible to improve the skill and specificity of their simulations and to generate potentially useful seasonal climate forecasts. Hidden Markov Models (HMM) have proven useful for capturing hidden states ('wet' or 'dry') and reproducing daily weather patterns and rainfall statistics such as the length of dry and wet spells that are important for crop choice and agricultural decisions (Greene et al., 2008). The addition of exogenous climate factors (Non-Homogeneous Hidden Markov Models (NHMM)) obtained from GCMs has improved the skill of the local stochastic models by including inter-annual variability important for multi-year aquifer planning and decadal water planning. (Bellone et al., 2004; Hughes and Guttorp, 1994; Hughes et al., 1999; Robertson et al., 2007; Robertson et al., 2004). In these models, the choice of climate factors (or predictors) is critical for reproducing within season and interannual rainfall variability. The challenge is that the GCM forecast fields are high dimensional with numerous potential predictors, many of whom are highly correlated. The best way to combine or select from these predictors in the NHMM is unclear. Previous studies have used meteorological indices derived from the GCM fields, such as mean sea level pressure and gradients of geopotential height across the region of interest (Hughes et al., 1999), as well as the GCM's regional precipitation field whose dimension is first reduced by principal component analysis

(PCA) (Robertson et al., 2004). Here, we present the use of a Support Vector Machine as a nonlinear alternative to PCA to generate a single input variable for the NHMM using multivariate reanalysis climate data.

2. STUDY AREA AND DATA

We apply the SVM-NHMM approach we develop to the eastern part of the Jordan River Valley. For Israelis and Palestinians, population density and per capita water usage has increased dramatically over the last century, and water availability and allocation is a source of tension. Eighty five percent of the water supply is rain-dependent with high intra-annual, inter-annual and decadal variability which in the past have been responsible for multi-year droughts. Palestinian agriculture is predominantly rain-fed, while in Israel it is largely irrigated using lake and ground water. Given its ability to store water in the aquifers, Israelis are more impacted by multi-year drought. Palestinians, however, are impacted by both intra-annual, or seasonal, as well as inter-annual variability. The impact of this variability on agricultural production, groundwater recharge, groundwater salinity, and lake level decline is of concern.

This region is located at a node between the influence of eastern Mediterranean and Middle Eastern climate regimes. Winter rainfall dominates and comes from the west with moisture advected from the Mediterranean. Mediterranean winter cyclones and mean-sea-level pressure (MSLP) variations have been shown to have monthly to yearly impacts on rainfall in Israel (Kutiel and Paz, 1998). Anomalously dry conditions in the region were found to be accompanied by positive MSLP anomalies in the eastern Mediterranean, and vice versa. Rainfall is thus highly influenced by the behavior of

Mediterranean storms, which have been found to explain up to 70% of the variance of rainfall in the region over the past 4 decades. (Trigo et al., 2000). (Alpert et al., 2004b) identify four main types of synoptic system in the Eastern Mediterranean (EM), each one dominant at a different time of year and bringing distinct weather patterns to the EM and surrounding region. The four main systems defined by Alpert et al (2004) include: (1) The Sharav lows, which come from the south-west primarily in the spring, bringing with them hot and dusty air from the Sahara or Arabian deserts; (2) Persian Troughs, which dominate in the summer, originating from the east and persisting along with the Asian Monsoon season. They cause relatively warm and humid air over the coastal EM region; (3) The Red Sea Troughs, prevalent in autumn/winter, originating from the south, they are associated with dry desert air. They are often pushed back by stronger winter systems, specifically Cyprus Lows; and (4) Mediterranean winter lows which come from the west bringing most of the rainfall over the continental part of the EM region. They are called Cyprus Lows when situated close to Cyprus. These lows often propagate from west to east over the Mediterranean, providing energy and moisture for cyclone development and causing Mediterranean storms along the way. These Mediterranean storms, or cyclones, often produce extreme weather events, with heavy rainfall.

2.1 Precipitation Data

Daily precipitation data from thirteen stations provided by the Israeli Meteorological Service were used. Figure 1 shows the locations of the selected stations. The period 1950–2000 was chosen because it has a complete record at all 13 stations, and   for which atmospheric reanalysis data is available. A list of the station names, locations, and source of the rainfall time series is given in Table 1.

Figure 2 shows box plots of the annual monthly average at the 13 stations. Most of the rainfall in the region occurs in the winter season between the months of October and April. Spatially, there is a sharp north-south precipitation gradient with the upper part of the Galilee and the Golan Heights receiving up to 1200 mm annually with around 600 mm annual average, while the Southern desert part receives about 50 mm annually.

2.2 Climate Data

Daily climate data over the Eastern Mediterranean (EM) region was selected from the 25 grid-boxes within the region (30 °E–40 °E, 27.5 °N–37.5 °N), with a resolution of 2.5° (Figure 3), obtained from the NCEP–NCAR Reanalysis Project (Kalnay et al. 1996; http://www.cdc.noaa.gov/cdc/reanalysis/ reanalysis.shtml). This domain captures the region of the Cyprus Lows and the Mediterranean storm track responsible for most of the rain in the region. For each gridbox, geopotential height, u-wind, and v-wind at 1000hPa were extracted as climate "predictors" of Jordon-valley rainfall (75 variables total). These variables have been used in other studies to automatically classify EM synoptic systems (Alpert et al., 2004b) as well as to determine the prevalence and evolution of different synoptic systems over the year and how they correlate to the different seasons – fall, winter, summer and spring (Alpert et al., 2004a). Here, we use reanalysis data to demonstrate the potential of the modeling approach. These same predictor fields can be extracted from GCMs making this approach potentially useful for future GCM downscaling studies.

3. METHOD

Figure 4 shows a flowchart of the methodology. There are three distinct parts. First, a hidden Markov model (HMM) without an exogenous predictor is run to determine the number of hidden states that are synoptically meaningful for the rainfall data set. The Viterbi sequence, which estimates the most likely sequence of hidden states that underlie the observed rainfall evolution is then identified. Next, two alternatives for dimension reduction of the reanalysis data are compared, namely principal components analysis (PCA) and the SVM. The PCA is applied directly to the atmospheric fields, while the SVM is applied to relate the atmospheric fields to the Viterbi sequence, in a manner analogous to a regression, or to Canonical Correlation Analysis. The SVM uses the reanalysis fields as inputs and the Viterbi sequence from the HMM as classifiers. A non-homogeneous HMM (NHMM) is then run using the predictors obtained from the two methods of dimension reduction, in order to predict the hidden state transition probabilities. The remainder of this section describes the SVM, HMM and NHMM components. The PCA is standard methodology, and was applied to the covariance matrix of the 75-variable GCM daily predictor field. For both the PCA and SVM, the GCM data was standardized at the outset by subtracting the long-term mean of each variable, and dividing by its long-term daily standard deviation.

3.1 Hidden Markov Models

A Hidden Markov Model considers a Markovian process to generate simulations of a given time series based on random sampling of the probability distribution functions (pdf's) conditioned on different hidden states, S, which are typically persistent synoptic weather regimes in our context. Figure 5 shows how the hidden weather states, S, are used to determine the output vectors of daily precipitation occurrences R. The model used is fully described in Robertson et al. (2004, 2007). In brief, the time sequence of daily rainfall measurements on a network of stations is assumed to be generated by a first-order Markov chain of a few discrete hidden (i.e. unobserved) rainfall states. For each state, the daily rainfall amount at each station is modeled as a finite mixture of components, consisting of a delta function at zero amount to model dry days, and a combination of two exponentials to describe rainfall amounts on days with non-zero rainfall. A maximum likelihood approach is used to estimate these rainfall “emission” parameters, as well as the transition matrix between states. Once these parameters have been estimated, the most-likely temporal sequence of states can be estimated given the rainfall data, by means of the Viterbi algorithm (Forney., 1973) this sequence is then used in conjunction with the reanalysis data to train an SVM.

The HMM is capable of generating variability directly based on the Markovian dynamics of the hidden variable S. However, it is necessary to incorporate exogenous variables in order to simulate inter-annual variability influenced by climate phenomena. In the NHMM, the state-transition matrix is treated as a (logistic) function of a multivariate predictor input time series (Figure 6), as described in Hughes and Guttorp (1994) and Robertson et al. (2004).

3.2 Support Vector Machine

The Support Vector Machine (SVM) is employed to generate a univariate daily time series from the 75-component reanalysis data. The SVM is a learning machine for classification problems and provides a way to classify multi-dimensional data based on the similarity of certain "attributes", learning by example to assign class-labels to objects so as to minimize both the prediction error and model complexity simultaneously (Noble, 2006). (Mukherjee and Mukherjee, 2002) showed that the SVM algorithm has a remarkable prediction capacity and it performed better than polynomial and rational approximations, local polynomial techniques, radial basis functions, and feed-forward artificial neural networks (ANNs) when applied on a database of chaotic time series. Unlike ANNs, the structure of the SVM model is not fixed in advance with a specific number of adjustable parameters, but can adapt to the data. Introduced by (Vapnik, 1995), the basic idea behind SVMs is to map the high-dimensional input space into a low dimensional *feature space* utilizing kernels. This so-called "kernel-trick" enables the SVM to work with feature spaces having very high dimensions. This makes it possible to perform the separation between different classes or groups even if they have very complex boundaries (See Figure 7)

Consider an example of binary classification (e.g., the Viterbi sequence for a two state HMM). One would like there to exist a w and b such that an observation x_i has "a positive class" if $w \cdot x_i - b >= 1$ and "negative class" if $w \cdot x_i - b \leq -1$. In order to find the plane that is furthest from both sets one has to maximize the margins between the support planes. The supporting graphs are maximally separated and supported by what are called support vectors. The margin between the two supporting planes is $\gamma = 2 / \| w^2 \|$, thus maximizing the margin is equivalent to:

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 \qquad (1)$$
$$\text{such that} \quad y_i(w \cdot x_i - b) \geq 1$$

Typically, the optimal parameters of Equation (1) are found by solving its dual formulation. After introducing a dual set of variables to construct a Lagrange function, and applying Karush-Kuhn-Tucker conditions, Vapnik (1995) has shown that Equation (1) is equivalent to the following in the dual form:

$$\min_{\alpha^*,\alpha} J_D\left(\alpha^*,\alpha\right) = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i)\cdot(\mathbf{x}_j) - \sum_{i=1}^{M}\alpha_i$$
$$\text{such that} \quad \sum_{i=1}^{M} y_i\alpha_i = 0 \qquad (2)$$
$$\alpha_i \geq 0, \forall_i$$

where the Lagrange multipliers α_i and α_j are required to be greater than or equal to zero for i = 1, …, M. Equation (2) comprises a convex constrained quadratic programming problem (Cortes and Vapnik, 1995; Vapnik, 1995). As a result, the *m-input vectors* that correspond to nonzero Lagrangian multipliers, α_i and α_j, correspond to the *support vectors*. The SVM model thus formulated, then, is guaranteed to have a global and unique solution.

Despite the mathematical simplicity and elegance of SVM training, it is able to identify relationships of high complexity (Liong and Sivapragasam, 2002; Scholkopf et al., 1997). The mapping of the data from the nonlinearly separable space to the linear space is carried out using feature functions $\Phi(\cdot)$; these functions attempt to perform mapping that is necessary for applying the linear algebra in the SVM formulation. The mapping to the feature space is carried out implicitly providing that the dot product of these mapping functions is equivalent to a well defined kernel functions that follow Mercer conditions (Vapnik, 1995). That is from Equation (2) the nonlinearity and the mapping to a space where linear algebra applies (i.e., the kernel trick) is computed as: $(\mathbf{x}_i)\cdot(\mathbf{x}_j) \rightarrow \Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j) \rightarrow K(\mathbf{x}_i,\mathbf{x}_j)$. The extension of SVM to multi-class classification is carried out by solving many binary classification SVMs and performing voting schemes afterward. Interested readers are referred to (Angulo and Catala, 2000; Angulo et al., 2003). Application of SVM in water resources management and chaotic time series analysis have been highlighted in (Khalil et al., 2005a; Khalil et al., 2005b; Khalil et al., 2006). Here, the Viterbi sequence from the HMM is what is classified, and the vector **x** corresponds to the atmospheric circulation variables considered as the predictors for the NHMM.

4. RESULTS AND ANALYSIS

4.1 HMM Results

The HMM was first implemented without climate predictors to determine whether the model could identify physically meaningful hidden states and capture persistence and trends. Based on log-likelihood, AIC and BIC scores, as well as an examination of the synoptic circulation patterns identified for the different states, the number of hidden states was set equal to four. The state transition probability matrix is presented in Table 2, and the Viterbi state-sequence plotted in Figure 8, with the states ordered from the driest to the wettest; overall rainfall attributes of each state are also given in Table 2. As can be seen in both the figure and the table the dry state (state ‘1’) and the wettest state (state

‘4’) are most persistent, with few direct transitions between them, with the mildly wet states 2 and 3 as transient intermediaries.

The meteorological characteristics associated with each state are shown in Figures 9 and 10, by making composites of reanalysis data for the days classified into each state by the Viterbi sequence. Figure 9 shows composites of sea level pressure anomalies (deviation from the long term monthly mean). Figure 10 shows composites of total wind at 850mb.

State 1 (the dominant state – 62% of the days) corresponds to dry conditions, with very weak rainfall intensities and very few rainfall occurrences. Winds from the west are very weak and an anomalous ridge overlays the region.

State 4 the ‘wettest’ state, is typical of a deep Cyprus or winter low, with a low pressure system just to the west of Cyprus, and strong moist westerly winds from the Mediterranean. Rainfall intensities are large, and 72% of the total rainfall occurs in this state.

States 2 and 3 can be described as transitions states. States 3 is similar to state 4 for the northern stations, with similar probability but lower intensity. Figure 9c shows a similarly structured but weaker MSLP anomaly. However, the main differences between states 3 and 4 are evident in the central and southern stations which have a much lower probability of rainfall in state 3 (41-48 % depending on the station) than in state 4 (71-99% depending on the station). While state 2 has both lower frequency and intensity rainfall than both states 3 and 4, the probability of rainfall in the central and south is similar to that in the north (25-53% for all stations). Figure 9b shows that for these days, a shallow Eastern Low is dominant, bringing low intensity rainfall to the entire country.

4.2 SVM and PCA Results

The SVM is next used in conjunction with the Viterbi sequence to generate a single climate predictor from the 75-component reanalysis data, thus using the HMM as the classifier, essentially mapping each day into a specific hidden state. Forty nine winter seasons (October–March, 182 days) between the years 1950 and 1999 were used (8918 days total). A subset of the training years was chosen as the computational costs of the model prohibited the use of the entire data set. The training set consisted of 4 years, or 728 days of data. Increasing the size of the training set to 8 years did not improve the results. The dates chosen for the training set included 4 seasons: 1956/1957, 1962/1963, 1968/1969, and 1969/1970. 1962/1963 was a drought year, 1956/1957 and 1969/1970 were average years with respect to rainfall amounts and 1968/1969 was a wet year. These seasons were selected to encompass the largest possible range of rainfall conditions. In the SVM, a radial basis function kernel was used (Hsu and Lin, 2002). Once the model was trained, it was tested on the entire 8918 day data set.

Table 3(a) presents a confusion matrix of the actual Viterbi sequence with the sequence predicted by the SVM, showing the number of times the predicted state (rows) matched the observed one (columns). Table 3(b) shows the percentage of days correctly and incorrectly classified for each state. The diagonal shows the days where the predicted class number matched the real class number. The other squares indicate how the model got ‘confused’, predicting class *y* when the actual class number was *x*. For example, there are 4386 days which were classed correctly as ‘1’ in the predicted set (the dominant dry state). In the square below, there were 552 days that were classed as ‘2’, but were actually class ‘1’. Even though the SVM misclassified about 35% of the days, it was still able to capture the within season and interannual rainfall variability.

4.3 A Comparison of NHMM Results using SVM and PCA

The NHMM was trained using the observed rainfall data, together with the SVM-predicted Viterbi sequence (Sect 4.2), both for the 29-year training period 1950–1979. The Viterbi sequence of 4 integer values was transformed into 4 binary timeseries, each representing a single state; these 4 binary predictors were then used as input to the NHMM without normalization. The combined SVM-NHMM was then validated using the independent 20-year period 1980–1999, using the reanalysis data as inputs, and verifying against the daily rainfall data. We compared the results generated when using the results from the SVM as the conditioning climate input to the results generated when using the leading 3 principal components (PCs) as input predictors. In this case, the PCA-NHMM was trained on the same 1950–1979 period using the PCs defined on that period as a 3-component input vector. For the 1980–1999 validation period, the reanalysis data was projected onto the PC loading vectors (ie the empirical orthogonal functions, EOFs) derived from the training period in order to derive the PCs as input to the NHMM during the validation period.

We ran 50 simulations for the test years and extracted seasonal rainfall statistics including the seasonal average and the number of wet and dry days, which are important for aquifer management, as well as the number of 7-day dry spells and the length of the maximum dry spell, which are important for agriculture management (crop choice and crop investment). The rainfall threshold here was defined at a day with more than 0.02 mm recorded rainfall. Figure 11 shows the averages (over the 13 stations and 50 simulations) of selected statistics along with the 50% and 95% confidence limits compared to the observed historical data for the PCA (panels a–c) vs SVM (panels d–f) models. The model using the SVM generated climate predictor performed better for all statistics analyzed. These include the average annual rainfall amount, 3-day consecutive rain events, 7-day consecutive dry spells, persistence statistics including wet-to-wet day and dry-to-dry day probabilities and the total number of wet days and dry days per season (Table 4). The SVM-NHMM model had higher correlations and lower mean squared errors and hence proved more skillful than the PC-NHMM model.

Predicted vs. observed rainfall at selected stations for the SVM-NHMM model is shown in Figures 12 and 13 for seasonal precipitation amounts and the number of 7-day dry spells respectively. For seasonal averages, all of the stations are captured well by the model, with the exception of Eilat. This is perhaps because the precipitation reaching Eilat is often associated with a Red Sea Trough, a system originating to the south in the Arabian peninsula which is often dry and not captured by the 4-state HMM . Precipitation to the other stations is dominated by the Cyprus Lows, the synoptic system captured in the climate data used to create the input vector. The stations in the south, in the more arid region, exhibit somewhat lower anomaly correlation scores than those in the wetter, northern region. This could perhaps be explained by the large variability and low total rainfall amount in those arid stations as well as by the fact that most of the variability (and rainfall) in the south is brought in by the Red Sea Trough. Alternatively, it could be due to pooling of all the results without scaling the data. Additional analysis of the 2 regions separately is currently in progress.

5. CONCLUSION

Statistical downscaling of precipitation from atmospheric GCMs involves choices of predictors, basis functions, and parameters as well as model form. The NHMM based strategies are somewhat more complex than simple regression models. However, they have been shown to be effective in past applications and are perhaps the most effective way to map the seasonal gridded forecasts to station daily precipitation scenarios for a season. One of the limitations in this technology has been the ability to effectively model potentially nonlinear relationships between the atmospheric circulation patterns and the parameters of a daily weather generator. Parsimony is an important goal in this regard as well. The experiment conducted in this paper establishes that the SVM-NHMM based approach can offer improved out-of-sample performance in terms of some key daily statistics of seasonal and inter-annual rainfall variability.

While reanalysis data was used in this study, fields of the same type are generated by GCMs and could be extracted for downscaling of seasonal climate forecasts, or as part of an evaluation of future climate scenarios. Some ad hoc choices for some of the model structure were made, including the number of training years for the SVM and the number of PCs used, but an effort was made to keep the design as consistent as possible across the PCA and SVM based designs.

The SVM classification used to develop the SVM-NHMM scheme is effectively a nonlinear encoding of the atmospheric circulation fields whereby a high-dimensional (here 75) vector describing the large-scale circulation field is assigned a state label (here derived from the station rainfall via the HMM). The resulting SVM-NHMM simulations exhibit stronger correlations with rainfall for stations in the northern region. This is probably partly because climatological rainfall is greater in the north and thus tends to play a larger role in determining the HMM states. Dynamical control of rainfall by circulation is also stronger in the north, associated with Cyprus Lows, while the Red Sea Trough that plays a role in rainfall in the south is more subtle and often accompanied by dry conditions (Tsvieli and Zangvil, 2005). Further analysis could potentially improve these results, perhaps by treating northern and southern stations separately, or by expanding the climate domain used as a predictor set. However the SVM-NHMM is still superior in performance to that of the corresponding PCA-NHMM using 3 PCs (that accounted for 63% of the variance of the predictor field). Thus, applications of the SVM-NHMM may be useful to explore as alternatives to the PCA-NHMM. As presented and applied here, the SVM-NHMM is essentially just as automatic a procedure as the PCA-NHMM, and cross-validated performance measures as deployed here can be used to choose between these two alternatives as well as different choices of the structure or classes for the classification variable. Additional experimentation with the method to assess its comparative performance is needed.

6. ACKNOWLEDGEMENTS

This work was supported in part by the US Department of Defense NDSEG Fellowship. NCEP–NCAR Reanalysis data were provided by the NOAA–CIRES Climate Diagnostics Center, Boulder, Colorado, from their Web site available online at http://www.cdc.noaa.gov. The rainfall data were provided by the Tel Aviv University Nasa node and the Israeli Meteorological Service as part of the GLOWA Jordan River project. The HMM code was developed by Sergey Kirshner and Padhraic Smyth, and can be obtained online at http://www.datalab.uci.edu/mvnhmm/. The Royal Netherlands Meteorological Institute (KNMI) database (www.climexp.knmi.nl) was developed and maintained by Geert Jan van Oldenborgh.

7. LIST OF FIGURES AND TABLES

Figure 13: Number of 7-day dry spells per season of SVM-NHMM-simulated rainfall amount for selected stations. (northern stations in the top row, central stations in the middle row and southern stations in the bottom row)The average of the 50 simulations is plotted for each year (dashed) together with the observed (solid). The number of dry spells (in days) per season is plotted on the ordinate.

Table 1: Station ID, name, source, longitude and latitude of selected stations

Table 2: The first four columns show the transition probabilities from each hidden state to hidden state for the four state HMM. The right-hand columns show the range of daily station average rainfall amount. For each state and the % of total rainfall that occurs during each state during the historical period 1950-1999.

Table 3: Confusion Matrix for the SVM, giving (a) the number and (b) the percentage of days that are classified correctly and incorrectly into each rainfall state for the 1950-1999 period. The observed states from the Viterbi sequence are in the columns and the predicted states from the SVM are in the rows. Correct classifications for each state can be seen along the diagonal.

Table 4: Comparison of Correlations and Mean Square Error (MSE) between SVM and PCA model with real data for specific rainfall statistics. Wet and Dry Spell Counts are the annual number of 3-day consecutive rain days and 7-day consecutive dry periods, respectively. The ratio of the SVM/PC MSE is less than 1, indicating the that SVM model fit the data better than the PC model.

8. REFERENCES

P. Alpert, I. Osetinsky, B. Ziv and H. Shafir, A New Seasons Definition based on Classified Daily Synoptic Systems: an Example for the Eastern Mediterranean, *International Journal of Climatology* **24**(2004a), pp. 1013-1022.

P. Alpert, I. Osetinsky, B. Ziv and H. Shafir, Semi-Objective Classification for Daily Synoptic Systems: Application to the Eastern Mediterranean *International Journal of Climatology* **24**(2004b), pp. 1001-1012.

C. Angulo and A. Catala, K-SVCR. A multi-class support vector machine, *Machine Learning: Ecml 2000* **1810**(2000), pp. 31-38.

C. Angulo, X. Parra and A. Catala, K-SVCR. A support vector machine for multi-class classification, *Neurocomputing* **55**(2003), pp. 57-77.

E. Bellone, J.P. Hughes and P. Guttorp, A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *NRCSE Technical Report Series No. 021*, National Research Center for Statistics and Environment, University of Washington (2004).

C. Cortes and V. Vapnik, Support-Vector Networks, *Machine Learning* **20**(1995), pp. 273-297.

G. Forney., The viterbi algorithm, *Proceedings of the IEEE* **61**(1973), pp. 268--278.

K.R. Gabriel and J. Neumann, A Markov chain model for daily rainfall occurence, *Quarterly Journal of the Royal Meteorological Society* **88**(1962), pp. 90-95.

L. Goddard *et al.*, Current approaches to seasonal to interannual climate predictions, *International Journal of Climatology* **21**(2001), pp. 1111-1152.

A.M. Greene, A.W. Robertson and S. Kirshner, Analysis of Indian monsoon daily rainfall on subseasonal to multidecadal time scales using a hidden Markov model, *Quarterly Journal of the Royal Meteorological Society* **134**(2008), pp. 875-887.

C.-W. Hsu and C.-J. Lin, A Simple Decomposition Method for Support Vector Machines, *Machine learning* **46**(2002), p. 291 (224 pages).

J.P. Hughes and P. Guttorp, Incorporating Spatial Dependence and Atmospheric Data in a Model of Precipitation, *Journal of applied meteorology / 33, no* **12**(1994), p. 1503.

J.P. Hughes, P. Guttorp and S.P. Charles, A non-homogeneous hidden Markov model for precipitation occurrence, *Applied statistics* **48**(1999), p. 15 (16 pages).

A. Khalil, M.N. Almasri, M. McKee and J.J. Kaluarachchi, Applicability of statistical learning algorithms in groundwater quality modeling, *Water Resources Research* **41**(2005a), pp. -.

A. Khalil, M. Mckee, M. Kemblowski and T. Asefa, Sparse Bayesian learning machine for real-time management of reservoir releases, *Water Resources Research* **41**(2005b), pp. -.

A.F. Khalil, M. McKee, M. Kemblowski, T. Asefa and L. Bastidas, Multiobjective analysis of chaotic dynamic systems with sparse learning machines, *Advances in Water Resources* **29**(2006), pp. 72-88.

H. Kutiel and S. Paz, Sea Level Pressure Departures in the Mediterranean and their Relationship with Monthly Rainfall Conditions in Israel, *Theoretical and applied climatology* **60**(1998), p. 93 (18 pages).

S.Y. Liong and C. Sivapragasam, Flood stage forecasting with support vector machines, *Journal of the American Water Resources Association* **38**(2002), pp. 173-186.

S.J. Mason, Simulating Climate over Western North America Using Stochastic Weather Generators, *Climatic Change* **62**(2004), pp. 155-187.

N. Mukherjee and S. Mukherjee, Predicting signal peptides with support vector machines, *Pattern Recogniton with Support Vector Machines, Proceedings* **2388**(2002), pp. 1-7.

W.S. Noble, What is a Support Vector Machine?, *Nature Biotechnology* **24**(2006).

C.W. Richardson, Stochastic Simulation of Daily Precipitation, Temperature, and Solar Radiation, *Water Resour. Res.* **17**(1981), pp. 182-190.

A.W. Robertson, A.V.M. Ines and J.W. Hansen, Downscaling of Seasonal Precipitation for Crop Simulation, *Journal of Applied Meteorology and Climatology*(2007), pp. 677-693.

A.W. Robertson, S. Kirshner, P. Smyth, Downscaling of Daily Rainfall Occurrence over Northeast Brazil Using a Hidden Markov Model, *Journal of Climate 17, no* **22**(2004), pp. 4407-4424.

B. Scholkopf *et al.*, Comparing support vector machines with Gaussian kernels to radial basis function classifiers, *Ieee Transactions on Signal Processing* **45**(1997), pp. 2758-2765.

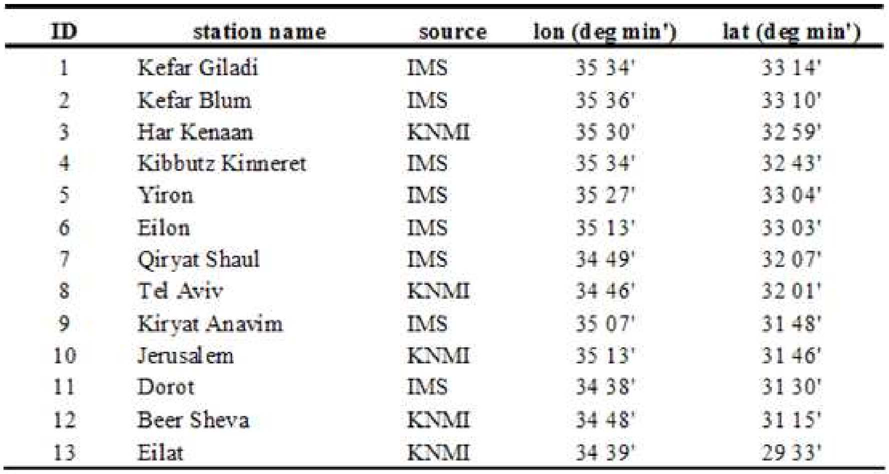
L. Sun, D.F. Moncunill, H. Li, A.D. Moura and F. A. S. Filho, Climate Downscaling over Nordeste Brazil using NCEP RSM97, *Journal of Climate* **18**(2005), pp. 551-567.

I.F. Trigo, T.D. Davies and G.R. Bigg, Decline in Mediterranean rainfall caused by weakening of Mediterranean cyclones *Geophysical research letters.* **27**(2000), p. 2913 (2914 pages).

Y. Tsvieli and A. Zangvil, Synoptic climatatological analysis of 'wet and 'dry' Red Sea Troughs over Israel, *International Journal of Climatology* **25** (2005), pp. 1997–2015.

V. Vapnik, *The Nature of Statistical Learning Theory* Springer-Verlag, New York, NY (1995).

Table 1: Station ID, name, source, longitude and latitude of selected stations.



ID	station name	source	lon (deg min')	lat (deg min')
1	Kefar Giladi	IMS	35 34'	33 14'
2	Kefar Blum	IMS	35 36'	33 10'
3	Har Kenaan	KNMI	35 30'	32 59'
4	Kibbutz Kinneret	IMS	35 34'	32 43'
5	Yiron	IMS	35 27'	33 04'
6	Eilon	IMS	35 13'	33 03'
7	Qiryat Shaul	IMS	34 49'	32 07'
8	Tel Aviv	KNMI	34 46'	32 01'
9	Kiryat Anavim	IMS	35 07'	31 48'
10	Jerusalem	KNMI	35 13'	31 46'
11	Dorot	IMS	34 38'	31 30'
12	Beer Sheva	KNMI	34 48'	31 15'
13	Eilat	KNMI	34 39'	29 33'

Table 2: The first four columns show the transition probabilities from each hidden state to hidden state for the four state HMM. The right-hand columns show the range of daily station average rainfall amount. For each state and the % of total rainfall that occurs during each state during the historical period 1950-1999.

		to state				Rainfall Amount (mm/day)	% of Total Rainfall
		1	**2**	**3**	**4**		
from state	1	0.79	0.1	0.08	0.03	0-3	0%
	2	0.49	0.24	0.17	0.1	0-6	5%
	3	0.24	0.28	0.26	0.22	1-13	22%
	4	0.05	0.21	0.27	0.47	6-50	73%

Table 3: Confusion Matrix for the SVM, giving the percentage of days that are classified correctly and incorrectly into each rainfall state for the 1950-1999 period. The observed states from the Viterbi sequence are in the columns and the predicted states from the SVM are in the rows. Correct classifications for each state can be seen along the diagonal. The far right column shows the number of days predicted for each state while the bottom rows shows the number of observed days for each state.

		Observed				
		1	2	3	4	Total
Predicted	1	84%	52%	34%	15%	5710
	2	11%	31%	21%	12%	1387
	3	4%	10%	24%	15%	805
	4	1%	7%	21%	58%	1016
	Total	5195	1441	1240	1042	**8918**

Table 4: Comparison of Correlations and Mean Square Error (MSE) between SVM and PCA model with real data for specific rainfall statistics over the validation period 1980-2000. Wet and Dry Spell Counts are the annual number of 3-day consecutive rain days and 7-day consecutive dry periods, respectively. The ratio of the SVM/PC MSE is less than 1, indicating the that SVM model fit the data better than the PC model.

	Correlation		**Mean Squared Error (MSE)**	
	PCA	**SVM**	**PCA**	**SVM/PCA ratio**
Amounts	0.62	0.81	226.53	0.65
Wet Spell Count	0.65	0.88	5.38	0.86
Dry Spell Count	0.57	0.85	4.01	0.36
Number Wet Days	0.58	0.86	19.17	0.58
Number Dry Days	0.58	0.86	19.17	0.58
Wet-to-Wet Probability	0.64	0.88	0.09	0.80
Dry-to-Dry Probability	0.53	0.83	0.12	0.44

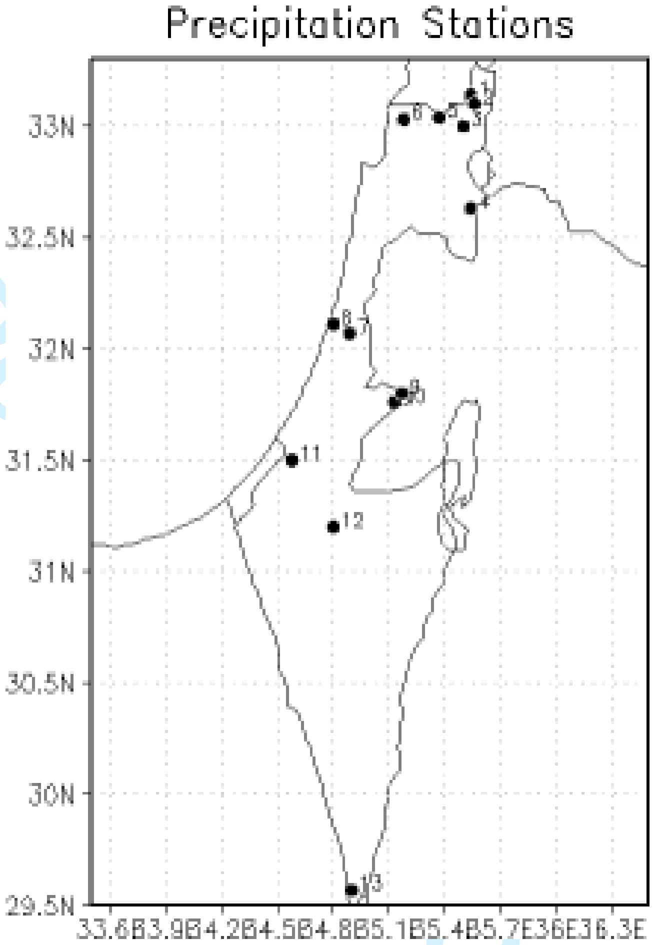


Figure 1: Map of 13 selected stations. Map generated using the KNMI Climate Explorer website.

.

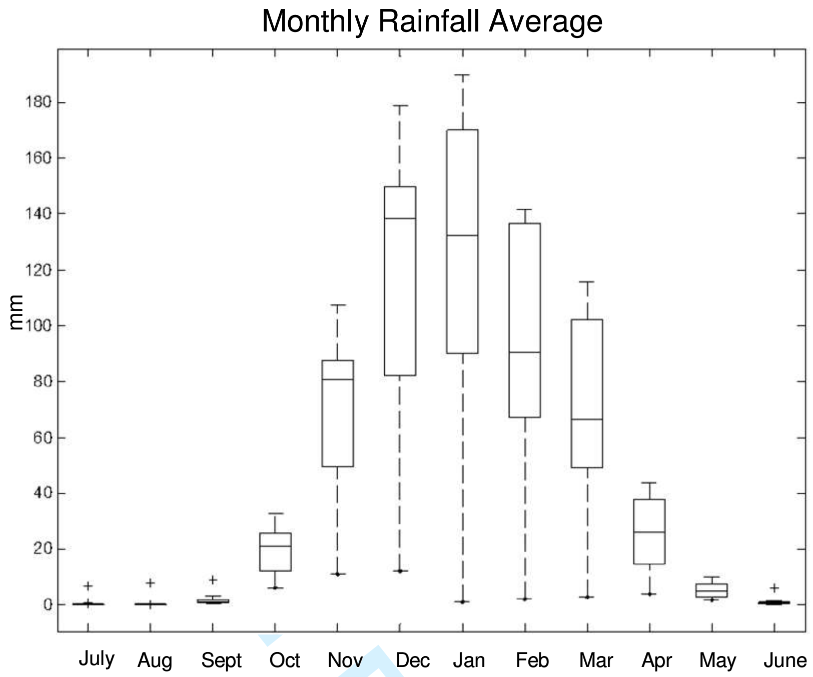


Figure 2: Boxplot of average monthly rainfall for all stations 1950-1999

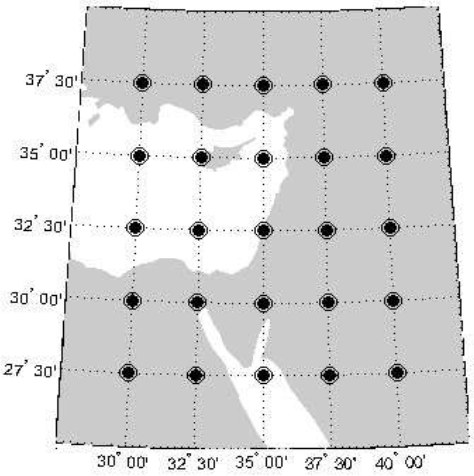


Figure 3: The 25 grid points from which geopotential height, u-wind, and v-wind are taken for climate predictors.

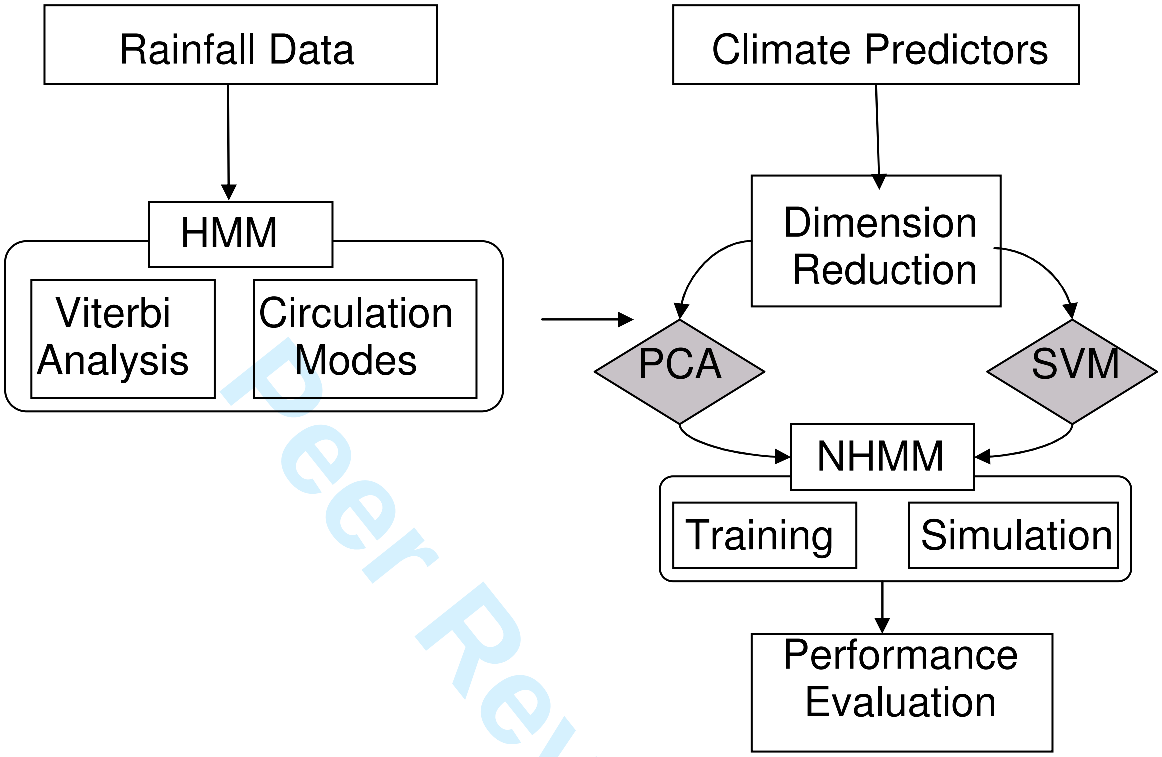


Figure 4: Flow Chart for model methodology.

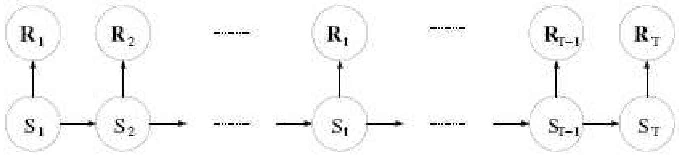


Figure 5: Graphical Model Representation of a Hidden Markov Model (Robertson et al., 2004).

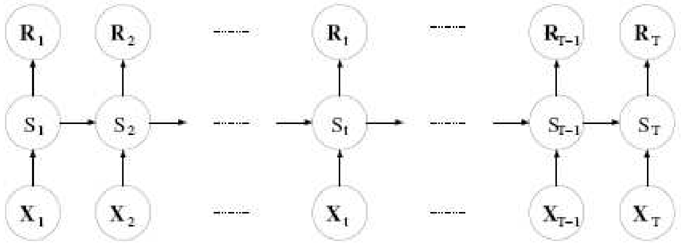


Figure 6: Graphical Model Representation of a Non-Homogeneous HMM (Robertson et al., 2004).

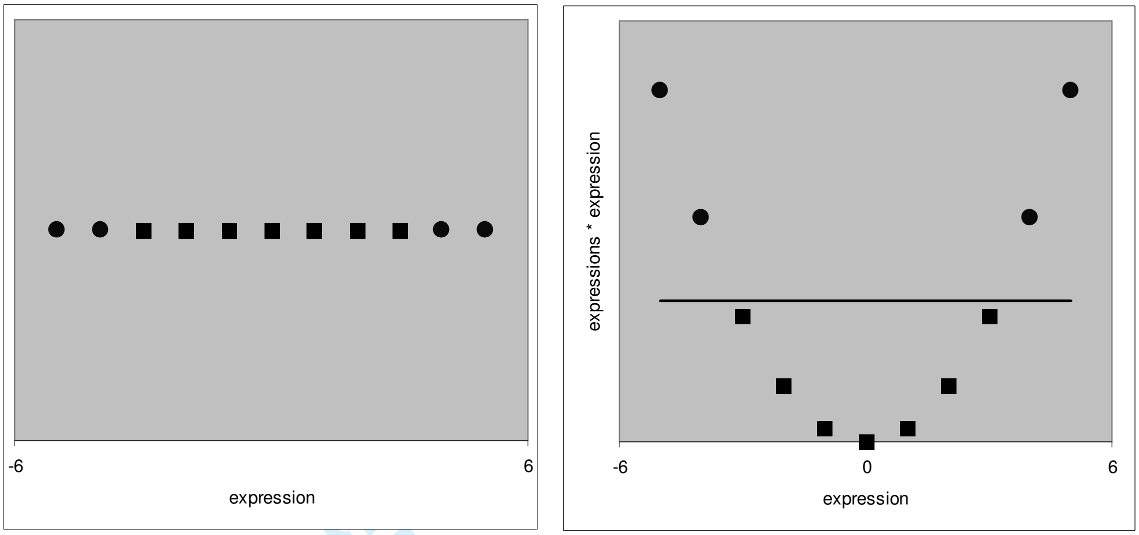


Figure 7: Conceptual representation of the kernel transformation to a higher dimensional feature space. A non-separable one-dimensional data set (left) multiplied by itself and transformed to a 2-dimensional space, where it is separable (right). Adapted from (Noble, 2006)

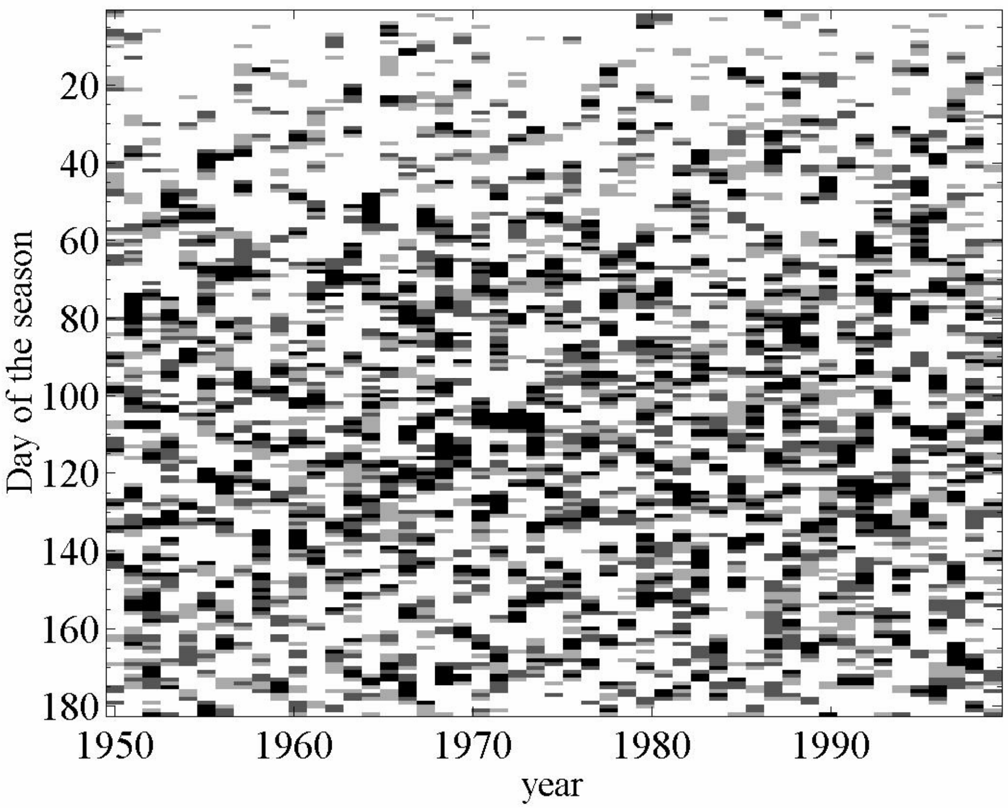


Figure 8: Estimated HMM state sequence of the historical rainfall data set from 1950-1999. State 1 is represented with white bars, state 2 with light grey bars, state 3 with dark grey bars and state 4 with black bars.

Figure 9: Geopotential Height anomaly composites for the four HMM states at 1000mb. For each day in the composite, anomalies are calculated as difference from the long term monthly mean.

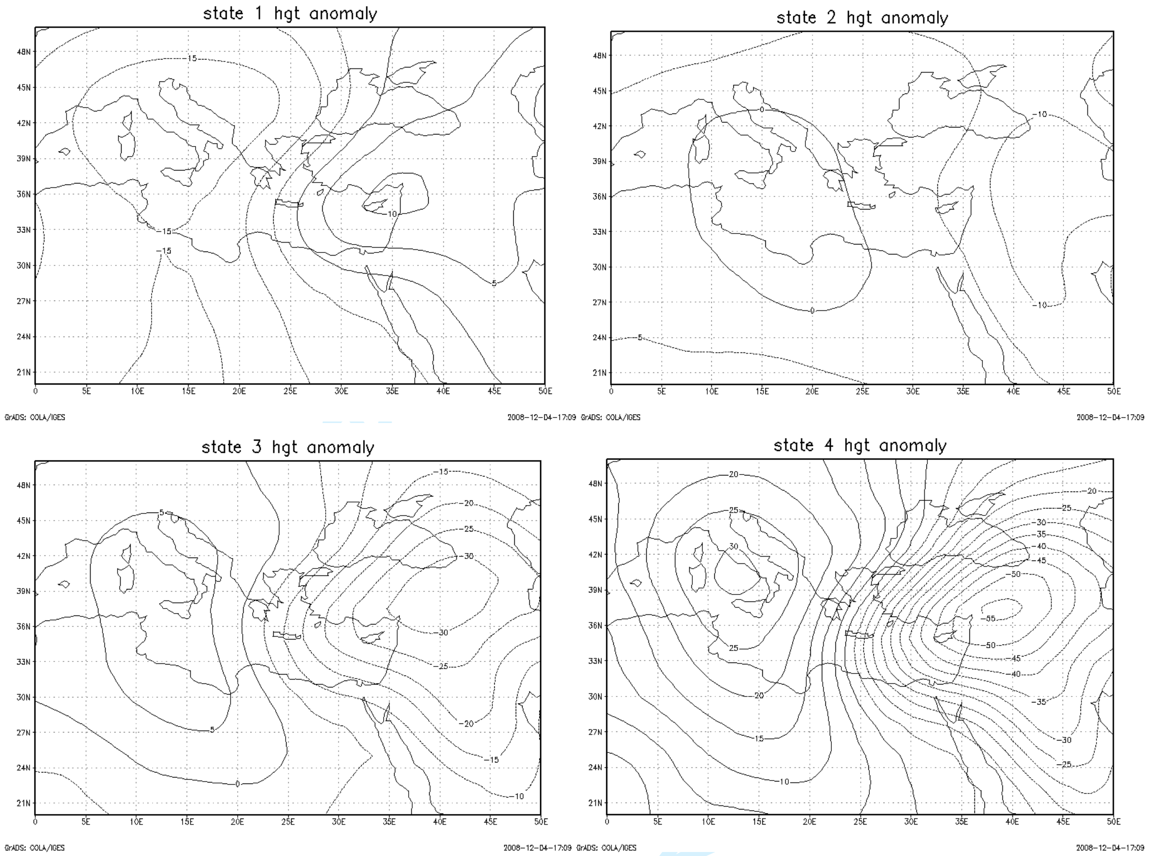
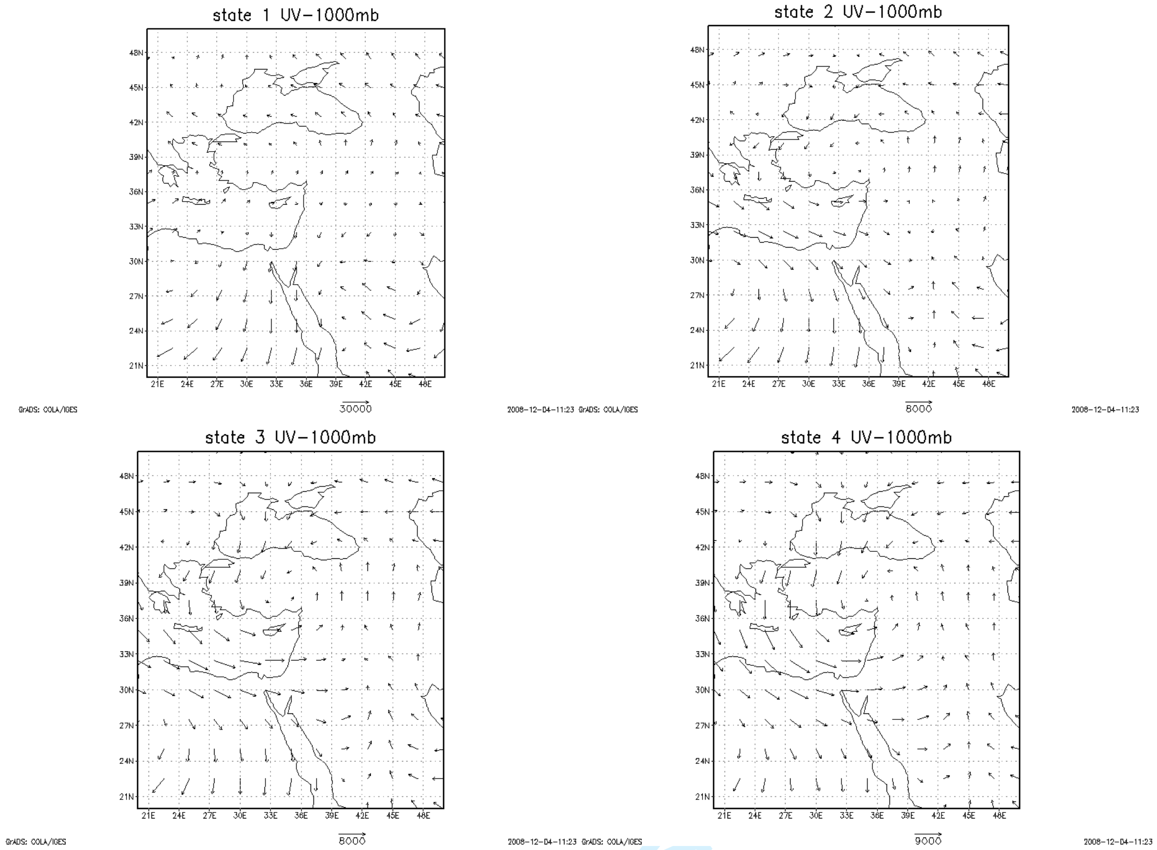


Figure 10: Total wind composites for the four difference states at 1000mb



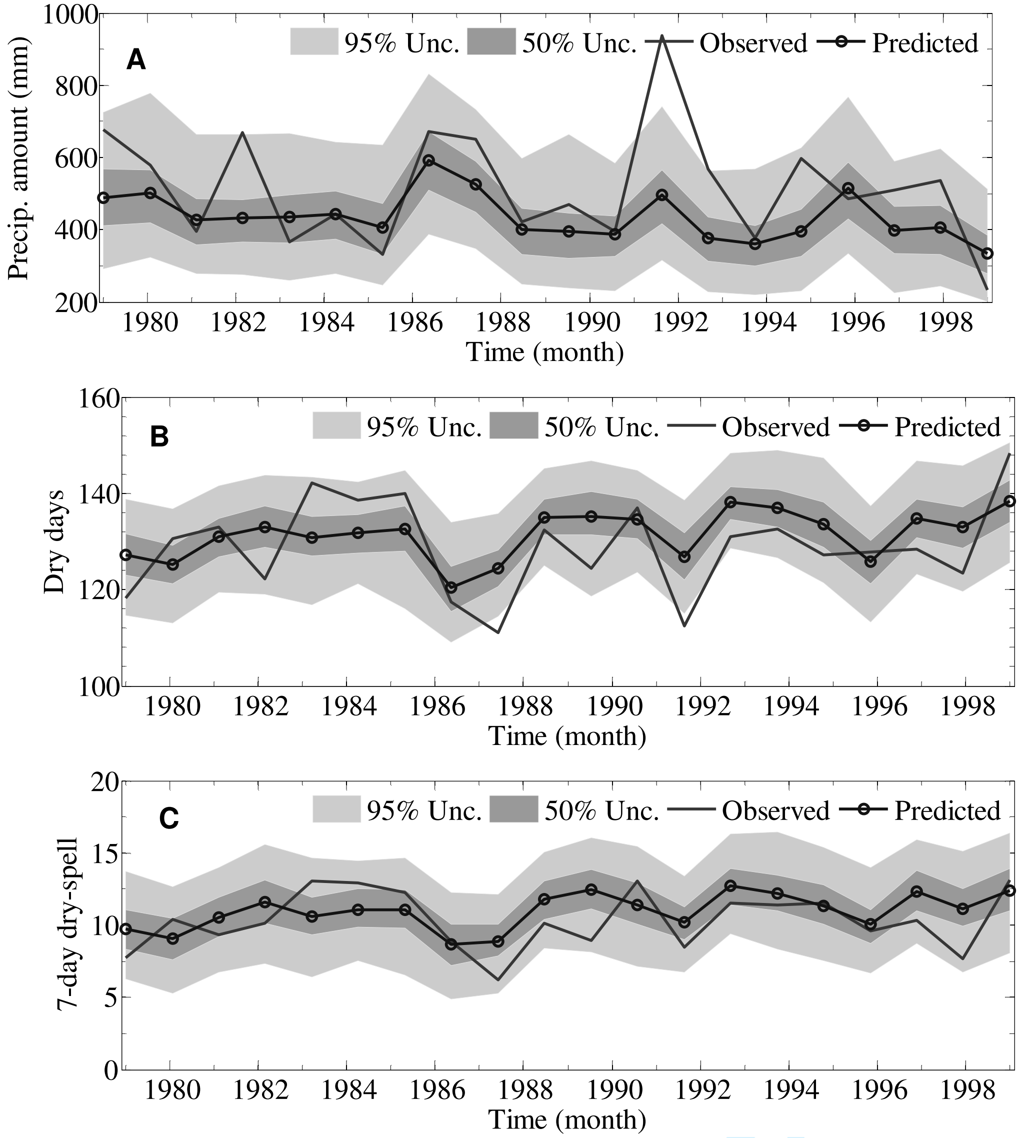


Figure 11 (a-c): Results from 50 NHMM simulations over the period withheld from model fitting. PCA generated climate predictors: Precipitation amounts, number of 7-day dry spells, and number of dry days per season averaged across all 13 stations. Solid lines are observed data and dotted lines are simulation averages. 50% and 95% confidence limits are shown.

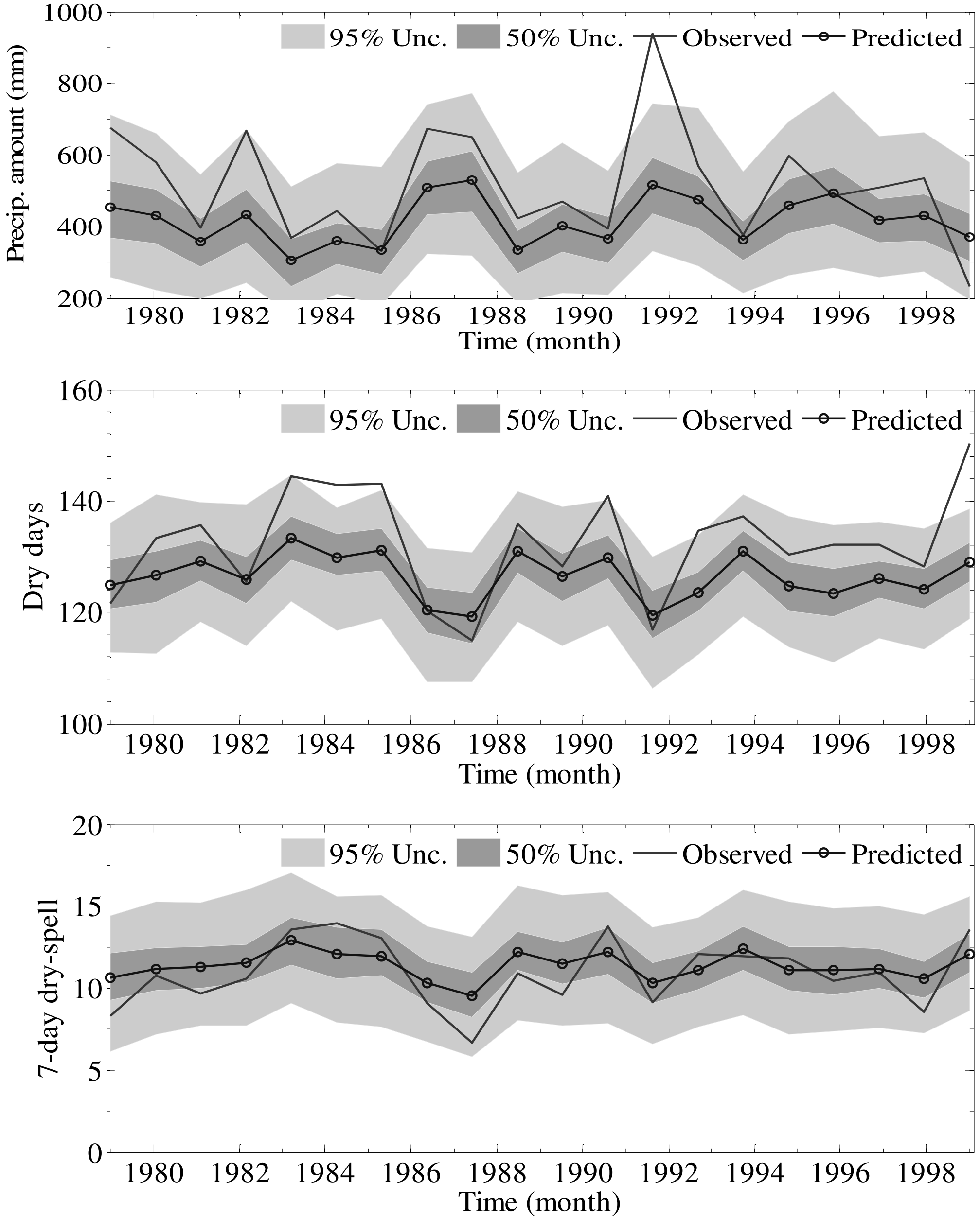


Figure 11 cont (d-f): Results from 50 NHMM simulations over the period withheld from model fitting. SVM generated climate predictors: Precipitation amounts, number of 7-day dry spells, and number of dry days per season averaged across all 13 stations. Solid lines are observed data and dotted lines are simulation averages. 50% and 95% confidence limits are shown. SVM results are better than PC results.

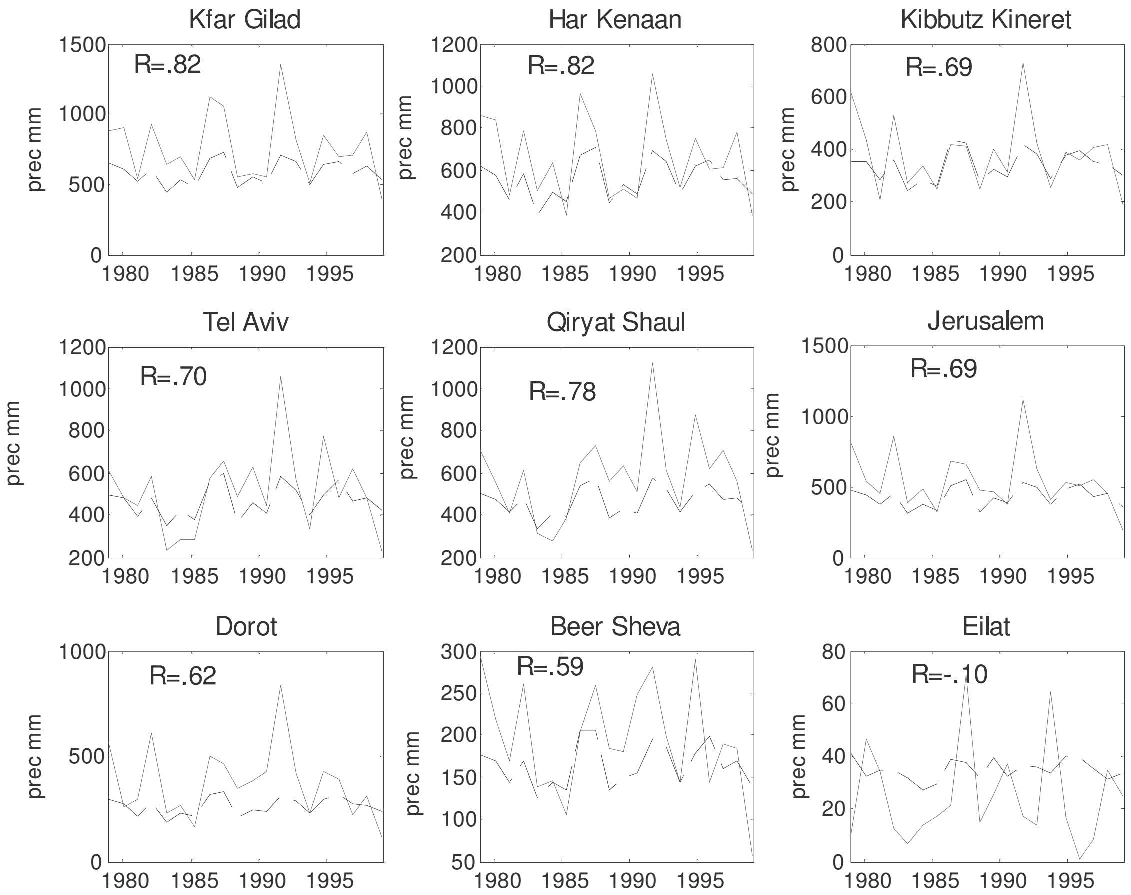


Figure 12: Interannual variability of SVM/NHMM-simulated rainfall amount for selected stations (northern stations in the top row, central stations in the middle row and southern stations in the bottom row). The average of the 50 simulations is plotted for each year (dashed) together with the observed (solid). The precipitation (in mm) per season is plotted on the ordinate. The results are only shown for years that were withheld from model fitting.

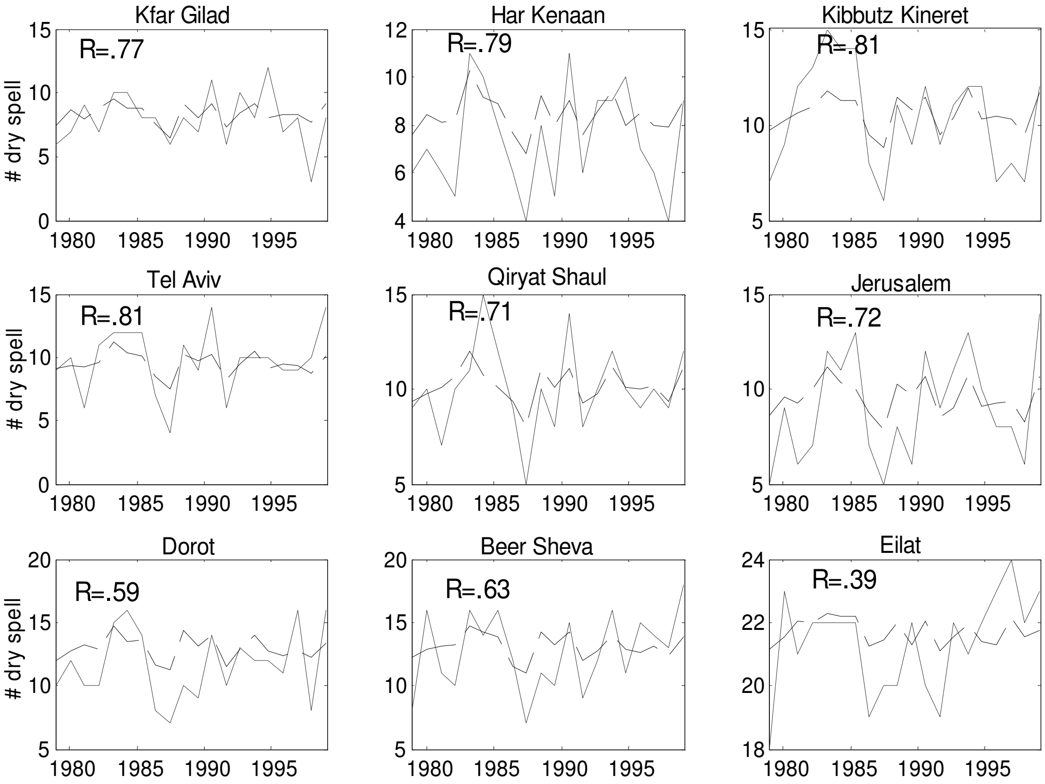


Figure 13: Number of 7-day dry spells per season of SVM/NHMM-simulated rainfall amount for selected stations. (northern stations in the top row, central stations in the middle row and southern stations in the bottom row)The average of the 50 simulations is plotted for each year (dashed) together with the observed (solid). The number of dry spells (in days) per season is plotted on the ordinate.