Downscaling atmospheric patterns to multi-site precipitation amounts in southern Scandinavia
Emiliano Gelati, Ole Bøssing Christensen, Peter F. Rasmussen and Dan Rosbjerg

ABSTRACT
A non-homogeneous hidden Markov model (NHMM) is applied for downscaling atmospheric synoptic patterns to winter multi-site daily precipitation amounts. The implemented NHMM assumes precipitation to be conditional on a hidden weather state that follows a Markov chain, whose transition probabilities depend on current atmospheric information. The gridded atmospheric fields are summarized through the singular value decomposition (SVD) technique. SVD is applied to geopotential height and relative humidity at several pressure levels, to identify their principal spatial patterns co-varying with precipitation. We assume the common hidden weather state process to completely account for the temporal structure of precipitation. Given the current weather state, the multivariate probability distribution of precipitation occurrences is approximated using a Chow–Liu tree dependence structure, involving products of bivariate distributions. Conditional on the weather state, precipitation amounts are modelled separately at each gauge as independent gamma-distributed random variables. This modelling approach is applied to 51 precipitation gauges in Denmark and southern Sweden for the period 1981–2003. The downscaling model produces robust predictions of data statistics, such as expected precipitation amounts and spell duration distributions. Moreover, the model-defined weather states show a satisfactory degree of physical consistency.

Key words | Chow–Liu trees, downscaling, hidden Markov model, precipitation model, southern Scandinavia

NOMENCLATURE AND ABBREVIATIONS

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>°</td>
<td>Arc degrees</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian information criterion</td>
</tr>
<tr>
<td>CO₂</td>
<td>Carbon dioxide</td>
</tr>
<tr>
<td>DMI</td>
<td>Danish Meteorological Institute</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation-maximization</td>
</tr>
<tr>
<td>GCM</td>
<td>General circulation model</td>
</tr>
<tr>
<td>GH-1000</td>
<td>Geopotential height at 1000 hectopascals</td>
</tr>
<tr>
<td>hPa</td>
<td>Hectopascals</td>
</tr>
<tr>
<td>km</td>
<td>Kilometers</td>
</tr>
<tr>
<td>LAM</td>
<td>Limited area meteorological model</td>
</tr>
<tr>
<td>m</td>
<td>Metres</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum likelihood estimate</td>
</tr>
<tr>
<td>NHMM</td>
<td>Non-homogeneous hidden Markov model</td>
</tr>
<tr>
<td>SMHI</td>
<td>Swedish Meteorological and Hydrological Institute</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular value decomposition</td>
</tr>
<tr>
<td>WSM</td>
<td>Weather state model</td>
</tr>
</tbody>
</table>

LIST OF SYMBOLS

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Matrix whose element $a_{ig}$ represents the covariance between an atmospheric field at node $i$ and the standardized precipitation process at gauge $g$</td>
</tr>
</tbody>
</table>

mm Millimetres

doi: 10.2166/nh.2010.114
c Threshold below which precipitation is neglected (fixed at 0.2 mm)

\(D\) Record length (total number of daily time steps)

\(\text{deg}(g)\) Number of Chow–Liu tree edges connecting gauge \(g\)

\(f^j_t(\cdot)\) Conditional precipitation amount pattern probability distribution assuming conditional spatial independence

\(f^T_t(\cdot)\) Tree approximation of the conditional multivariate distribution of the precipitation occurrence pattern

\(f^R_t(\cdot)\) Tree approximation of the conditional multivariate distribution of the precipitation amount pattern

\(E_j\) Chow–Liu tree defined for state \(j\)

\(F\) Number of free model parameters

\(G\) Number of precipitation gauges

\(Ga\left(\kappa_{gj}, \varphi_{gj}\right)\) Two-parameter gamma probability density function with parameters \(\kappa_{gj}\) and \(\varphi_{gj}\)

\(L(\cdot)\) Model likelihood function

\(M, U, W\) Matrices resulting from the singular value decomposition of \(A\)

\(o_{gj}\) Precipitation occurrence probability at gauge \(g\), if state \(j\) prevails

\(P_{ij}\) Base-line transition probability of shifting from state \(i\) to state \(j\)

\(q_{gh(1.1)|j}\) Probability of precipitation to occur at both gauges \(g\) and \(h\), conditioned on state \(j\)

\(q_{gh(1.0)|j}\) Probability of precipitation to occur at gauge \(g\) and not at gauge \(h\), conditionally on state \(j\)

\(q_{gh(0.1)|j}\) Probability of precipitation to occur at gauge \(h\) and not at gauge \(g\), conditionally on state \(j\)

\(Q(\cdot)\) Conditional expectation of model log-likelihood

\(r_{gt}\) Precipitation amount at time step \(t\) at gauge \(g\)

\(r_t\) Precipitation amount pattern at time \(t\)

\(R^t_{t_1}\) Precipitation amount patterns from time step \(t_1\) to \(t_2\)

\(s_t\) Prevailing weather state at time step \(t\)

\(S\) Number of model-defined weather states

\(S^t_{t_1}\) State sequence from time \(t_1\) to \(t_2\)

\(S^D_1\) Most likely state sequence identified by the Viterbi algorithm

\(T_{gh}(y_{gt}, y_{ht}|s_t = j)\) Bivariate distribution of precipitation occurrences at gauges \(g\) and \(h\), conditional on state \(j\)

\(T_g(y_{gt}|s_t = j, \theta)\) Marginal distribution of \(T_{gh}(y_{gt}, y_{ht}|s_t = j, \theta)\) with respect to \(y_{ht}\), for any \(h \neq g\)

\(u_j\) Elements of the diagonal matrix \(U\)

\(V\) Covariance matrix of the atmospheric variables

\(x_t\) Values of atmospheric variables at time step \(t\)

\(X^{t_2}_{t_1}\) Values of atmospheric variables from time step \(t_1\) to \(t_2\)

\(y_{gt}\) Precipitation occurrence value at gauge \(g\) at time \(t\)

\(y_t\) Precipitation occurrence pattern at gauge \(g\) at time \(t\)

\(z\) Generic multi-dimensional standardised atmospheric field

\(\alpha(f), \beta(f)\) Forward and backward variables (Baum–Welch algorithm) of state \(j\) at time step \(t\)

\(\kappa_{gj}, \varphi_{gj}\) Gamma distribution parameters for gauge \(g\), if state \(j\) prevails

\(\mu_{ij}\) Expected values of the atmospheric variables corresponding to a shift from state \(i\) to \(j\)

\(\theta\) Model parameter set

\(\theta^{(n)}\) Parameter estimates at the \(n\)th iteration of the expectation-maximization algorithm

\(\omega_j(f)\) Highest probability of ending in state \(j\) at time \(t\)

\(\psi_j(f)\) Most likely state to occur at time \(t - 1\), given that state \(j\) occurs at time \(t\)

**INTRODUCTION**

Predictions of global climate change have stimulated an interest in assessing hydrologic effects at regional scales. Precipitation is a key factor in the description of hydrologic regimes and a required input to hydrological models. Consequently, there has been a renewed interest in stochastic weather generation to be used in climate change studies (Hughes et al. 1993).
Traditionally, stochastic approaches have not included atmospheric information as a conditioning factor for precipitation simulation. Stern & Coe (1984) described precipitation occurrences with a non-stationary Markov chain, but although model parameters were seasonal, their variations were merely time-dependent and not forced by any climatic signal. Another example is Zucchini & Guttorp (1991) who developed a multi-site model where precipitation was described separately at each site, given a common unobserved climatic process.

It is widely recognized that precipitation models that do not include atmospheric data should not be applied under climatic conditions different from those used for calibration, as the processes driving precipitation are likely not to be stationary (Hughes et al. 1999). The formulation of non-homogeneous hidden Markov models (NHMMs) originates from the desire to generate time series that are consistent with atmospheric historical records and with outputs of general circulation models (GCM) (Hughes & Guttorp 1994).

GCMs are the main tool for producing climate change scenarios caused by increasing CO₂ concentrations (Hughes et al. 1995). As GCMs typically simulate climatic variables on very large grids, they are not suitable for predicting non-smooth fields such as precipitation (Hughes & Guttorp 1994). Previous studies, including Giorgi & Mearns (1991) and Bates et al. (1998), pointed out the need for models that can downscale GCM simulations and historical synoptic-scale atmospheric records to local-scale daily precipitation.

Giorgi & Mearns (1991) downscaled GCM predictions by nesting a finer mesh-limited-area meteorological model (LAM) in a GCM, with the latter providing the boundary conditions for the higher resolution modelling. However, this methodology involves high computational costs that may limit its practical use. Moreover, inaccuracies and biases of GCM simulations are likely to propagate to the local scale (Giorgi & Mearns 1991; Hughes et al. 1993).

Although traditional stochastic precipitation models did not include atmospheric information as a conditioning factor, the attractiveness of Markov chains for modelling precipitation has long been recognized. Stern & Coe (1984) traced it back to early attempts to fit distributions of wet/dry spell durations (Cooke 1953; Green 1970). Other authors (e.g. Haan et al. 1976) modelled precipitation with Markov chains shifting between several states, with each state representing a certain range of precipitation amounts.

Weather state models (WSMs) constitute a stochastic approach to downscaling where synoptic-scale atmospheric signals are translated into local precipitation patterns. The concept of WSM was developed by Hay et al. (1991) and is based on assigning observed days to weather types that can be defined through expert meteorological knowledge (Bardossy & Plate 1991) or automatic classification methods (Hughes et al. 1993). After such assignments, precipitation probability distributions are estimated for each weather state.

NHMMs for precipitation were developed by Hughes & Guttorp (1994), Hughes et al. (1999) and Bellone et al. (2000). As in the case of WSMs, NHMMs can simulate precipitation conditioned on atmospheric information. However, in NHMMs the weather types are described as states of a Markov chain and are not defined a priori by classifying atmospheric patterns. Instead, weather states are identified via model fitting where each weather state is characterized by a separate precipitation probability distribution, while the atmospheric variables influence state transitions. Thus the NHMM weather states are identified taking due account of both atmospheric and precipitation patterns. This may constitute an advantage in comparison with WSMs, where weather types result from an aprioristic classification of atmospheric information (Charles et al. 1999). Recently, Robertson et al. (2004) used an NHMM to downscale daily precipitation occurrence over northeastern Brazil. Robertson et al. (2007) used NHMM precipitation simulations over the south-eastern United States as input for a crop model.

In this study, NHMMs are applied to a dense network of precipitation gauges in Denmark and southern Sweden. To account for the spatial dependence structure of precipitation, we used Chow–Liu trees (Chow & Liu 1968; Meila & Jordan 2000) to model the probability distribution of precipitation occurrence patterns (i.e. at multiple gauges) within each weather state. Chow–Liu trees allow approximating multivariate distributions of discrete random variables using products of bivariate distributions. Kirshner et al. (2004) embedded Chow–Liu trees within
NHMMs in order to model multivariate precipitation occurrence time series in south-western Australia.

THE NON-HOMOGENEOUS HIDDEN MARKOV MODEL

Model formulation

NHMMs assume that precipitation is driven by an unobserved weather state process. The weather state at time $t$ is represented by the discrete stochastic variable $s_t$ that follows an $S$-state first-order Markov chain, where $S$ is the number of model-defined weather states. The transition probabilities for the Markov chain are assumed to depend on current atmospheric variables, making the Markov process non-homogeneous. At each time step $t$, the probability distribution of precipitation at one or more stations in a region is assumed to be uniquely determined by the weather state prevailing on that day.

Let $r$ denote a vector of length $G$ of precipitation amounts on day $t$ at $G$ stations in a region and let $x_t$ be a vector of $N$ atmospheric variables that are known to influence precipitation, for example measures of geopotential height and/or relative humidity. The probability of being in a particular weather state on day $t$ is determined by the weather state at time $t-1$ and by $x_t$. The parameters of the probability density function of $r$ depend on the state at time $t$. Thus the hidden weather state process entirely accounts for the temporal dependence structure of the observed precipitation process. This constitutes the assumption of conditional temporal independence of precipitation.

Weather state transition probabilities are assumed to be products of a baseline component corresponding to a stationary Markov chain transition probability and a function of the atmospheric information. Applying the Bayes theorem, we can deduce

$$
Pr[s_t=j|s_{t-1}=i, x_t] = \frac{\Pr[s_t=j|s_{t-1}=i] \Pr[x_t|s_{t-1}=i]}{Pr[s_{t-1}=i]} \propto \Pr[x_t|s_{t-1}=i] Pr[s_{t-1}=i]$$

(1)

The above transition probability function is parameterized according to Hughes & Guttorp (1994):

$$
Pr[s_t=j|s_{t-1}=i, x_t, \theta] \propto p_{ij} \exp \left[ -\frac{1}{2}(x_t - \mu_{ij})^T V^{-1}(x_t - \mu_{ij}) \right]
$$

(2)

where $V$ is the covariance matrix of the atmospheric variables; $p_{ij}$ is the baseline transition probability from state $i$ to $j$; $\mu_{ij}$ can be interpreted as the vector of expected values of the atmospheric variables when state $j$ occurs, given that state $i$ prevailed at the previous time step; and $\theta$ is the set of model parameters. Equation (2) involves the kernel of a multivariate normal distribution. This is a somewhat arbitrary choice and other kernels could be considered.

A gamma distribution is often used to model daily precipitation accumulation on wet days at a station. Let $Ga(r, \kappa_g, \varphi_g)$ denote a two-parameter gamma density with parameters $\kappa_g$ and $\varphi_g$ that describe the distribution of daily precipitation at station $g$ when the weather is in state $j$. If we assume conditional spatial independence of precipitation, i.e. daily precipitation amounts at different gauges are independent given the weather state, then a multivariate precipitation distribution can be obtained by multiplying individual densities. The density of daily precipitation at a single gauge can be defined as a mixture of a discrete probability of no precipitation and a gamma distribution. Thus, assuming conditional spatial independence, the conditional multivariate precipitation distribution can be expressed according to Bellone et al. (2000):

$$
f_r(r|s_t=j, \theta) = \prod_{g=1}^{G} \left[ o_{gj} Ga(r_{gt} - c; \kappa_g, \varphi_g) \right]^{y_{gt}} (1 - o_{gj})^{1-y_{gt}}
$$

(3)

where $r_{gt}$ is the precipitation amount on day $t$ at gauge $g$; $y_{gt}$ is the precipitation occurrence value on day $t$ at gauge $g$, which is equal to 1 if $r_{gt} \geq c$ and 0 otherwise; the threshold $c$ is set to 0.2 mm, which is the typical resolution of a tipping bucket rain gauge; $G$ is the number of rain gauges; and $o_{gj}$ is the precipitation occurrence probability at gauge $g$ for state $j$. Equation (3) relies on the assumption that precipitation processes at multiple gauges are independent, given the common weather state. However, as concluded by Hughes et al. (1999), the hypothesis of conditional spatial independence is not suitable for dense networks of gauges such as the one studied here.

To model explicitly the spatial correlation of precipitation occurrences, Kirchner et al. (2004) combined
NHMMs with the Chow–Liu tree model for multivariate discrete data (Chow & Liu 1968; Meila & Jordan 2000). The Chow–Liu method approximates a multivariate distribution of $G$ discrete variables using the product of $G-1$ bivariate distributions. The $G-1$ pairs of variables, which are chosen from among the possible $G(G-1)/2$ pairs, constitute the edges of the tree. Let $y_i$ denote the vector of precipitation occurrences on day $t$ at $G$ gauges. Let $f_j(y_i|s_t=j)$ be the multivariate distribution of $y_i$, conditioned on state $j$. Defining $E_t$ as the Chow–Liu tree associated with weather state $j$, the tree approximation of $f_j(y_i|s_t=j)$ is

$$f_j^T(y_i|s_t=j, E_t, \theta) = \frac{\prod_{(g,h) \in E_t} T_{gh}(y_{gt}, y_{ht}|s_t=j)}{\prod_{g=1}^G T_g(y_{gt}|s_t=j)^{\deg(g)-1}}$$

(4)

where $T_{gh}(y_{gt}, y_{ht}|s_t=j, \theta)$ is the bivariate distribution of precipitation occurrences at gauges $g$ and $h$, conditional on state $j$; $T_g(y_{gt}|s_t=j, \theta)$ is the marginal of $T_{gh}(y_{gt}, y_{ht}|s_t=j, \theta)$ with respect to $y_{ht}$ for any $h \neq g$; and $\deg(g)$ is the number of edges connecting gauge $g$. The distribution $T_{gh}$ is defined as

$$T_{gh}(y_{gt}, y_{ht}|s_t=j, \theta) = q_{gh}(y_{gt}|y_{ht}, s_t=j)q_{gh}(y_{ht}|y_{gt}, s_t=j)
\quad \times \left[1 - q_{gh}(y_{ht}, s_t=j) - q_{gh}(y_{ht}|y_{gt}, s_t=j) + q_{gh}(y_{ht}|y_{gt}, s_t=j)ight]^{1-\gamma_{gh}}$$

(5)

where $q_{gh}(y_{ht}|y_{gt}, s_t=j) = \Pr[y_{ht}=1, y_{gt}=1|s_t=j]$, $q_{gh}(y_{ht}|y_{gt}, s_t=j) = \Pr[y_{ht}=1, y_{gt}=0|s_t=j]$ and $q_{gh}(y_{ht}|y_{gt}, s_t=j) = \Pr[y_{ht}=0, y_{gt}=0|s_t=j]$. Consistently with Equations (3) and (5), the marginal distribution $T_g$ is defined as

$$T_g(y_{gt}|s_t=j, \theta) = \sigma_{g,2}^0(1-\sigma_{g,2}^0)^{1-\gamma_{gh}}$$

(6)

By embedding Chow–Liu trees to approximate the conditional (state-specific) precipitation occurrence pattern distributions, and modelling precipitation amounts separately at each gauge, the conditional multivariate precipitation distribution becomes

$$f_j^T(r_i|s_t=j, E_t, \theta)$$

$$= \frac{\prod_{(g,h) \in E_t} T_{gh}(y_{gt}, y_{ht}|s_t=j, \theta)}{\prod_{g=1}^G T_g(y_{gt}|s_t=j)^{\deg(g)-1}}\prod_{g=1}^G \Gamma(a_{g,2})$$

(7)

Thus the spatial correlation of precipitation is accounted for by the approximate multivariate tree distribution of precipitation occurrences, while amounts are assumed to be conditionally independent given the weather state and the precipitation occurrence dependence tree structure. In this study we adopted an NHMM using Equation (7) to model the precipitation conditional density (NHMM-T), while an NHMM using Equation (3) (NHMM-I) helped to benchmark the improvement in reproducing spatial correlation of precipitation due to Chow–Liu trees.

### Atmospheric information

The inclusion of atmospheric variables is a key step in developing an NHMM for precipitation. Atmospheric information is normally obtained from gridded reanalysis data. Given the high dimension of such datasets, a dimension reduction technique should be employed. A good summarization technique should reduce the gridded data to a few values and, at the same time, preserve as much as possible of the covariance between the atmospheric fields and precipitation.

In this application, we used singular value decomposition (SVD) to reduce the dimension of the atmospheric information. For each standardised atmospheric field $z_i$, one can define a matrix $A$ whose element $a_{ig}$ represents the covariance between the standardized atmospheric field at node $i$ and the standardized precipitation process at gauge $g$. The matrix $A$ can be decomposed by SVD:

$$A = MUW^T$$

(8)

where the columns of $M$ and $W$ are the eigenvectors with non-zero eigenvalues of, respectively, $AA^T$ and $A^TA$; and $U$ is a diagonal matrix, whose elements $u_i$ are the square roots of the non-zero eigenvalues of $AA^T$ (or equivalently $A^TA$) (Strang 1988). The summary variable $x_i$ (i.e. the $i$th element of $x$) is obtained by multiplying $z$ by the $i$th column of $M$. The variable $x_i$ explains the fraction $u_i^2/\sum_{i=1}^G u_i^2$ of covariance between $z$ and the precipitation process (Bretherton et al. 1992).

### Model estimation

Estimation of the NHMM described above is made complicated by the fact that weather states are not actually observed. In this study, maximum likelihood estimates
(MLEs) of model parameters are obtained using the iterative expectation-maximization (EM) algorithm (Hughes et al. 1999). The parameter set \( \theta \) to be estimated is composed of: the \( S(S - 1) \) baseline transition probabilities \( p_{ij} \); the NS\((S - 1) \) mean values of summary variables \( \mu_k \); the 2GS parameters of gamma distributions \( r_{ij} \) and \( \sigma_{ij} \); the GS gauge precipitation occurrence probabilities \( a_gj \); and the \( 3(G - 1)S \) edge precipitation occurrence pattern probabilities \( q_{gh(1,1)ij} \), \( q_{gh(1,0)ij} \) and \( q_{gh(0,1)ij} \). The parameter estimates at the \( m \)th iteration of the EM algorithm are denoted \( \hat{\theta}^{(m)} \). Each iteration of the EM algorithm consists of two steps. The first step (expectation) computes the expected value of the model log-likelihood function given the previous parameter estimates \( \hat{\theta}^{(m-1)} \):

\[
Q(\hat{\theta}^{(m-1)}, \hat{\theta}^{(m-1)}; \hat{\theta}^{(m-1)}) = E\left[ \log L(\hat{\theta}^{(m-1)}, X_1^D) | \hat{\theta}^{(m-1)} \right]
\]

\[
= \sum_{i=1}^{S} \sum_{j=1}^{D} \Pr[S_t = i, S_t = j | \hat{R}_{i1}^D, \hat{X}_1^D, \hat{\theta}^{(m-1)}] \times \log \Pr[S_t = j | S_{t-1} = i, X_t = x_t, \theta]
\]

\[
+ \sum_{j=1}^{S} \sum_{i=1}^{D} \Pr[S_t = j | \hat{R}_{i1}^D, \hat{X}_1^D, \hat{\theta}^{(m-1)}] \log f_r^T(r_t | S_t = j, E_j, \theta)
\]

where \( D \) is the record length; \( R_1^D \) and \( X_1^D \) are, respectively, the measured rainfall patterns and the atmospheric variables from time step 1 to \( D \); and the state probability terms can be computed by the Baum–Welch algorithm (Baum et al. 1970; Rabiner 1989; Akintug & Rasmussen 2005), described in some detail in Appendix A.

The second step (maximization) maximizes \( Q(\theta | \hat{\theta}^{(m-1)}) \) with respect to \( \theta \), leading to the new estimates

\[
\hat{\theta}^{(m)} = \arg\max \theta Q(\theta | \hat{\theta}^{(m-1)})
\]

As shown in Equation (10), the two terms summing to \( Q \) involve different parameters. Thus they can be maximized individually using Lagrange multipliers enforcing the constraints \( \sum_{i=1}^{S} P_{ij} = 1 \) and \( \sum_{j=1}^{S} \mu_i = 0 \) (for \( i = 1, \ldots, S \)).

However, the maximization of the second term requires choosing the edges constituting the Chow–Liu trees \( E_j \), for \( j = 1, \ldots, S \). At each iteration \( n \) of the EM algorithm, edge selections are carried out after the single-state occurrence probabilities \( \Pr[S_t = j | \hat{R}_{i1}^D, \hat{X}_1^D, \hat{\theta}^{(m-1)}] \) have been calculated via the Baum–Welch algorithm. For any weather state \( j \), we define the conditional mutual information between \( y_{gt} \) and \( y_{st} \) as

\[
I^{(n-1)}_{gh} = \sum_{i=1}^{D} T_{gh}(r_{gi}, r_{hi}| S_t = j, \hat{\theta}^{(m-1)}) \times \log \left( \frac{T_{gh}(r_{gi}, r_{hi}| S_t = j, \hat{\theta}^{(m-1)})}{T_{gh}(r_{gi}, r_{hi}| S_t = j, \hat{\theta}^{(m-1)})} \right)
\]

\[
\times \Pr[S_t = j | \hat{R}_{i1}^D, X_1^D, \hat{\theta}^{(m-1)}]
\]

\[
(11)
\]

The optimal tree approximation of \( f_r(y_t|S_t = j) \) is given by the tree \( E_j \) that maximizes the sum of conditional mutual information of the edges (Chow & Liu 1968; Kirshner et al. 2004):

\[
E_j^{(m)} = \arg\max \sum_{i=1}^{S} I_{gh}^{(m-1)}
\]

Chow & Liu (1968) proposed a method to find the optimal tree \( E_j^{(m)} \): (i) the possible \( G(G - 1)/2 \) edges are sorted according to decreasing values of conditional mutual information; (ii) the first two edges with the highest \( I_{gh}^{(m-1)} \) score are included in \( E_j^{(m)} \); (iii) the edge with the immediately lower \( I_{gh}^{(m-1)} \) score is selected if it does not form a circle with the already included edges; (iv) step (iii) is iterated until \( G - 1 \) edges are included. Once the optimal trees \( E_j^{(m)} \) are defined for \( j = 1, \ldots, S \), the new parameter estimates \( \hat{\theta}^{(m)} \) are calculated via the maximization step.

The EM steps are iterated until a maximum of the likelihood function is reached. However, the likelihood of an NHMM depends on a large number of parameters and may have several local maxima. Therefore the choice of initial parameter values is essential for obtaining proper estimates. To initialize the EM algorithm, we applied the convergent variant of MacQueen’s k-means clustering method (Anderberg 1973) to the \( G \)-dimensional time series \( R_1^D \) to have an initial guess of the hidden weather state sequence.

**Model selection**

Choosing the best model set-up implies deciding how many weather states are to be defined and which atmospheric covariates to be included. This means trading off between...
goodness-of-fit and parameter parsimony. We used the Bayesian information criterion (BIC) (Kass & Raftery 1995) as a performance indicator. The BIC appears to yield reasonable models, although some assumptions underlying its definition are violated by its application to NHMMs (Hughes et al. 1999). Defining $F$ as the number of free parameters, the BIC is

$$\text{BIC}(\hat{\theta}) = -2 \log L(\hat{\theta}; R_{11}, X_{1}^{P}) + F \log(GD)$$  \hspace{1cm} (13)

**STUDY AREA AND DATA**

The study area is southern Scandinavia, consisting of Denmark and southern Sweden (refer to Figure 1(a) for an overview of the study area). This region is located in the so-called western wind belt, in which weather is characterized by fronts, low pressure systems and changing conditions. The weather is strongly influenced by both the Atlantic Ocean and the continent, implying that the weather type changes according to the dominating wind direction (Danish Meteorological Institute 1997).

**Meteorological overview**

When western winds blow, low pressure systems move in from the northern Atlantic Ocean to Scandinavia and bring frontal precipitation. As lows often come in a row, this weather type can persist from some days up to a few weeks. If low pressure systems move eastwards at a safe distance from southern Scandinavia, high pressure weather prevails with stable and dry conditions. If winds blow westwards, the weather becomes continental. The eastern wind regime occurs when the high pressure area is broken over the continent but is still intact over Fennoscandia. Air masses, which are heated by the Baltic Sea, may bring precipitation over the eastern coast of Götaland and the Baltic islands. Although this weather type is very stable, eastern winds do not blow over southern Scandinavia as often as western winds. Like eastern winds, southern winds come from the continent. During summer, these air masses convey humidity, often provoking squalls or thunderstorms. North is the least frequent origin for winds blowing over southern Scandinavia. Due to the lee effect of the Norwegian mountains, north-western winds bring dry weather over northern Jutland, Zealand and Götaland, while precipitation may occur on southern and western Jutland. North-eastern winds often cause precipitation over Denmark, as the cold air masses coming from Sweden are heated by the Kattegat.

**Data**

Daily precipitation and atmospheric data were extracted for the period 1981–2003, from November to February. We considered only autumn–winter periods because of the recognized difficulties of associating spring and summer precipitation with synoptic-scale atmospheric patterns (Linderson 2001). NHMMs are indeed not likely to make good predictions for seasons and regions where convection
drives precipitation: convective processes are characterized by fine spatial scales and are not predictable by synoptic atmospheric information (Hughes et al. 1999).

Precipitation measurements were made available by DMI (Danish Meteorological Institute) and SMHI (Swedish Meteorological and Hydrological Institute). Data from 51 gauges were used and their locations are shown in Figure 1(b). Given the low frequency of missing observations (0.75%), we executed data-filling by multiple linear regression.

The atmospheric fields are NCEP Reanalysis Data provided by the NOAA-CIRES Climate Diagnostics Center (Boulder, CO, USA). Geopotential height and relative humidity, at several pressure levels, have been chosen because of their relevance to the precipitation phenomenon: geopotential height represents the circulation patterns, while relative humidity indicates the degree of saturation of the atmospheric layers. The atmospheric grid nodes (spaced 2.5° in latitude and longitude) used for this application are plotted in Figure 1(b).

RESULTS

The records were split into calibration (1981–1996) and validation (1997–2003) datasets. We chose the best model by comparing setups including two or three atmospheric covariates and 5–11 weather states. BIC scores indicated a group of best performing models, but its value had relatively
small variations among these set-ups. Hence, the final choice was made incorporating other considerations. We wanted to include at least one covariate accounting for synoptic circulation and one for air humidity, but we also wanted to avoid redundancies in atmospheric inputs and state definitions. Thus we chose an NHMM with 8 weather states, downscaling geopotential height at 1,000 hPa and relative humidity at 850 hPa.

**Precipitation statistics**

Here we present model-based reproductions of some precipitation statistics, for both calibration and validation periods. The simulated statistics are estimated by averaging over a thousand model runs, conditioning on the observed atmospheric time series. We chose the gauge at Tranebjerg (marked with a square in Figure 1(b)) to show the gauge-specific results, namely spell duration distributions and precipitation quantiles. The gauge at Tranebjerg was chosen as it well represents the average reproduction of gauge-specific statistics. As overall statistics, mean precipitation amounts and lag-0 cross correlations are reported in Figures 2 and 3.

![Observed mean precipitation amounts are reproduced with good precision for the calibration period](http://iwaponline.com/hr/article-pdf/41/3-4/193/370948/193.pdf)
while validation estimates are slightly less accurate (Figure 2(b)).

Lag-0 cross-correlations are measures of linear dependence between precipitation processes at pairs of gauges at the same time. Here we evaluate how the reproduction of such statistics is improved by embedding Chow–Liu trees (NHMM-T), with respect to assuming conditional spatial independence of precipitation (NHMM-I). Figure 3 shows the results for NHMM-I: high correlations are systematically underestimated by the model, for both calibration (Figure 3(a)) and validation (Figure 3(b)) periods. By means of a linear regression between cross-correlations and inter-gauge distances, we identified an approximate critical distance of 270 km, below which the hypothesis of

![Figure 3](image-url)

Figure 3 | The reproduction of dry- and wet-spell duration distributions at Tranebjerg gauge for (a) calibration (1981–1996) and (b) validation (1997–2003) periods.

![Figure 4](image-url)

Figure 4 | Weather state 1: (a) precipitation occurrence probabilities and edges of the optimal Chow–Liu tree, (b) expected precipitation amounts (mm) and (c) averaged geopotential height at 1,000 hPa.
conditional spatial independence is not appropriate. The fraction of couples of gauges below this distance is about 40%, thus embedding a spatial dependence model of precipitation seems necessary. Figure 4 shows the reproduction of lag-0 cross correlations for NHMM-T: for both calibration (Figure 4(a)) and validation (Figure 4(b)) periods, the reproduction of high correlations is generally improved. However, such estimates are still affected by underestimation, though to a smaller extent than for NHMM-I. NHMM-T embeds a spatial dependence model for precipitation occurrences, while amounts are modelled separately at different gauges given the weather state and the precipitation occurrence pattern. Thus, correlations between precipitation amounts at highly correlated gauges are likely to be underestimated by NHMM-T as well. A spatial dependence model for precipitation amounts may lead to further improvement.

Figure 5 plots observed versus simulated precipitation quantiles for the gauge at Tranebjerg. The good reproduction during both calibration (Figure 5(a)) and validation (Figure 5(b)) periods suggests that the gamma distribution is well suited for this application.

The reproduction of dry- and wet-spell duration distributions is satisfactory for both calibration (Figure 6(a)) and validation (Figure 6(b)). This suggests that the assumption of temporal conditional independence is acceptable for reproducing the generally low persistence of the precipitation process in this region. Indeed, auto-correlation coefficients (at any lag) are found to be lower than 0.2 for most gauges.

Weather states

The model-defined weather states allow further validation of the model by checking whether state-specific precipitation probability patterns are physically consistent with their corresponding averaged atmospheric patterns. The latter are obtained as weighted averages of the atmospheric gridded time series, where the weights are the state occurrence probabilities calculated by the Baum–Welch algorithm. Here we present precipitation occurrence probabilities, the defined Chow–Liu trees, expected amounts and averaged geopotential height at 1,000 hPa (GH-1000) for all eight weather states.
Weather state 1 is characterized by very small precipitation occurrence probabilities at all gauges (Figure 7(a)). Expected precipitation amounts are close to zero in the whole region (Figure 7(b)). The GH-1000 pattern depicts a strong high pressure area over continental Europe and southern Scandinavia (Figure 7(c)). Hence, state 1 corresponds to the typical high pressure weather.

State 2 corresponds to a westerly weather driven by a deep low pressure system approaching Scandinavia. Precipitation occurrence probabilities and expected amounts are large at all gauges (Figure 8(a)). At several gauges in Götaland, expected precipitation exceeds 10 mm (Figure 8(b)). GH-1000 values suggest that winds blow from the south west, causing precipitation in the whole region (Figure 8(c)).

For state 3, the GH-1000 pattern depicts a weak westerly regime over Scandinavia, where winds prevalently blow from the north west (Figure 9(c)). Precipitation occurrence probabilities are large in Denmark and low in Götaland. Probabilities are largest in western Jutland (Figure 9(a)). Precipitation is least likely to occur in northern Götaland, probably because of the lee effect due to the Norwegian mountains. Expected amounts have a similar distribution and are below 4 mm at all gauges (Figure 9(b)).

State 4 is characterized by large precipitation occurrence probabilities in Götaland, especially in the western part, by very low values in eastern Denmark, and by intermediate values in Jutland (Figure 10(a)). Such probabilities can correlate with the GH-1000 values if we hypothesize that atmospheric circulation corresponds to the early development of a low pressure system triggering a weak westerly regime (Figure 10(c)). In this case, air masses moving from the south west and being heated by the Kattegat are more likely to cause precipitation in Götaland than in Denmark. Figure 10(b) shows that expected precipitation amounts are smaller in Denmark (below 2 mm) than in Götaland (above 2 mm at nearly all gauges). In north-western Götaland expected amounts exceed 5 mm, while smaller values characterize the southern, central and eastern parts.

Precipitation occurrence probabilities of state 5 are largest in north-western Götaland and decrease towards the east and south, while they are small in all of Denmark (Figure 11(a)). Expected amounts are below 1 mm in the whole region except for north-western Götaland, where they exceed 2 mm at a few gauges (Figure 11(b)).
The GH-1000 pattern depicts an even less developed low pressure system than for state 4 (Figure 11(c)). Consistently, in state 5 precipitation is likely to occur in a smaller area than in state 4. Thus there is good physical agreement between the two states.

For state 6, precipitation occurrence probabilities are relatively large in the eastern-most part of the region and progressively decrease westwards (Figure 12(a)). Expected amounts are close to zero at all gauges, except for those located in eastern Götaland. However, at such gauges precipitation is not expected to exceed 2 mm (Figure 12(b)). GH-1000 values show a high pressure system over Scandinavia and Russia, corresponding to an easterly regime (Figure 12(c)). Under these conditions, air masses move from the east and are heated and moisturized by the Baltic Sea. Thus precipitation is most likely to occur in the eastern part of the region.

State 7 is characterized by a low pressure system over Norway, suggesting that winds blow from the west (Figure 13(c)). Precipitation occurrence probabilities are very large in Denmark and central Götaland (Figure 13(a)). The smaller probabilities in eastern and western Götaland may be due to intense precipitation falling on Denmark and on the hills of central Götaland, thus causing air masses to lose moisture. In Denmark, expected amounts are between 5 and 11 mm, with the largest values in the western part. In Götaland expected precipitation ranges from 1 to 2 mm in the eastern and western parts, and from 2 to 4 mm in the centre (Figure 13(b)).

The precipitation occurrence probability pattern defining state 8 is similar to state 7, but with lower values at all gauges except for those located in Jutland and central and north-western Götaland (Figure 14(a)). The largest precipitation amounts are expected in western Denmark and north-western Götaland (from 2 to 5 mm), while values below 2 mm characterize the rest of the region (Figure 14(b)). Generally expected amounts are also smaller than for state 7. Consistently, the GH-1000 values describe a westerly regime that is weaker than for state 7 (Figure 14(c)).

To summarize, states 1 and 6 describe high pressure conditions; states 3 and 5 depict either developing or extinguishing low pressure systems; states 4 and 8 may represent low pressure systems approaching Scandinavia; states 2 and 7 describe low pressure systems that have...
Figure 11  |  Weather state 5: (a) precipitation occurrence probabilities and edges of the optimal Chow–Liu tree, (b) expected precipitation amounts (mm) and (c) averaged geopotential height at 1,000 hPa.

Figure 12  |  Weather state 6: (a) precipitation occurrence probabilities and edges of the optimal Chow–Liu tree, (b) expected precipitation amounts (mm) and (c) averaged geopotential height at 1,000 hPa.
Figure 13 | Weather state 7: (a) precipitation occurrence probabilities and edges of the optimal Chow–Liu tree, (b) expected precipitation amounts (mm) and (c) averaged geopotential height at 1,000 hPa.

Figure 14 | Weather state 8: (a) precipitation occurrence probabilities and edges of the optimal Chow–Liu tree, (b) expected precipitation amounts (mm) and (c) averaged geopotential height at 1,000 hPa.
reached Scandinavia moving, respectively, northwards and eastwards. To have an overview of the most occurring weather state transitions we can apply the Viterbi algorithm (Viterbi 1967; Rabiner 1989). The Viterbi algorithm, described in Appendix B, is used to find the most likely hidden state sequence given observations and model parameters (Figure 15). Such analysis reveals that state 1 is most likely to persist. If not persisting, state 1 cannot evolve into state 2, 7 or 8. State 5 may evolve into state 1, 3, 4 or 8. State 4 is most likely to evolve into state 2, 7 or 8, but cannot shift to state 1. State 2 mostly shifts to state 7 and cannot shift to state 1, while state 7 is most likely to evolve into states 3 and 4. State 3 is most likely to evolve into states 1 and 4. State 8 cannot persist and is most likely to shift to state 3. State 6 is most likely to shift to state 1, but cannot evolve into state 2, 7 or 8. To conclude, most weather states representing low pressure systems at various stages (2, 3, 4, 7 and 8) are likely to shift among each other. State 5 might indicate an extinguishing low pressure system. Indeed, it is most likely to evolve into state 1 and, similarly to high pressure states (1 and 6), cannot shift to most wet states.

The edges of the Chow–Liu trees seem to constitute reasonable spatial correlation structures for all weather states. Indeed, most edges connect neighboring gauges, and only in a very few cases a dependence connection is characterized by a large distance.

CONCLUSIONS

The applied NHMM adequately predicts key statistics such as expected precipitation amounts and spell length duration distributions for both calibration and validation periods.

Validation results highlight the potential of NHMMs for producing realistic simulations under altered climate scenarios. However, the ability of NHMMs to downscale climate change scenarios cannot be inferred on the basis of the presented results. The good physical significance of the weather classes is encouraging in this aspect. Although no direct weather classifier was used to identify the states, the associated averaged atmospheric patterns showed reasonable correlations with the precipitation occurrence probabilities and expected amounts defining the weather states. Moreover, these correspondences provided insight into the atmospheric processes that drive precipitation over the region.

The physical consistency of the weather classes suggests that the SVD-based summarizing technique succeeded in capturing the principal co-varying spatial patterns of the atmospheric fields. This conclusion is also supported by the good predictions during validation.

NHMMs embedding Chow–Liu trees (NHMM-T), to approximate the multivariate probability distribution of precipitation occurrences, improve the reproduction of the observed spatial correlation of precipitation. This improvement is evaluated by benchmarking the performance of NHMM-T against NHMM-I, which assumes conditional spatial independence of the precipitation process. However, we believe that there is room for further improvement, as NHMM-T parameterizes explicitly spatial dependence for precipitation occurrences, but not for amount distributions.

REFERENCES

APPENDIX A. THE BAUM–WELCH ALGORITHM

To apply the EM algorithm, it is necessary to calculate the probability terms $\Pr[s_t = j|\mathbf{R}_t^D, \mathbf{X}_t^D, \hat{\theta}^{n-1}]$ and $\Pr[s_{t-1} = i, s_t = j|\mathbf{R}_t^D, \mathbf{X}_t^D, \theta]$. The Baum–Welch algorithm calculates these state probabilities from the following form:

\begin{align}
\Pr[s_t = j|\mathbf{R}_t^D, \mathbf{X}_t^D, \theta] & \propto \alpha_t(j)\beta_t(j) \\
\Pr[s_{t-1} = i, s_t = j|\mathbf{R}_t^D, \mathbf{X}_t^D, \theta] & \propto \alpha_{t-1}(i)\Pr[s_t = j|s_{t-1} = i, \mathbf{x}_t, \theta]P_{ij}F_i(r_t|s_t = j, E_j, \theta)\beta_t(j)
\end{align}

where $\alpha_t(j)$ and $\beta_t(j)$ are referred to as the forward and backward variables, respectively. $\alpha_t(j)$ is the probability of observing the sequence of precipitation patterns $\mathbf{R}_t^D$ when state $j$ occurs at time $t$ (Bengio 1999):

$$\alpha_t(j) = \Pr[\mathbf{R}_t^D, s_t = j|\mathbf{X}_t^D, \theta]$$

The probability terms $\alpha_t(j)$ are computed by forward recursion:

$$
\alpha_0(j) = \begin{cases} 1 & \text{if } j = E_0, \\
0 & \text{otherwise}
\end{cases}
$$

for $t = 1, 2, \ldots, T$,

$$
\alpha_t(j) = \sum_{i} \alpha_{t-1}(i)P_{ij}F_i(r_t|s_t = j, E_j, \theta)
$$

and $\beta_T(j) = 1$, for $t = T, T-1, \ldots, 1$,

$$
\beta_t(j) = \sum_{i} \beta_{t+1}(i)P_{ij}F_i(r_t|s_t = j, E_j, \theta)
$$

The Baum–Welch algorithm calculates these state probabilities from the following form:

$$
\Pr[s_t = j|\mathbf{R}_t^D, \mathbf{X}_t^D, \hat{\theta}^{n-1}] \propto \alpha_t(j)\beta_t(j)
$$

First received 1 December 2008; accepted in revised form 10 December 2009. Available online April 2010
\[ \alpha_t(j) \]
\[ \propto \begin{cases} f_t^T(r_t|s_t=j,E_t,\theta)\Pr[s_t=j|x_t,\theta], & \text{if } t=1 \\ f_t^T(r_t|s_t=j,E_t,\theta)\sum_{i=1}^{S} \alpha_{t-1}(i)\Pr[s_{t-1}=i|x_t,\theta] & \text{if } t>1 \end{cases} \]
\[ \beta_t(j) = \Pr[r_t^D|s_t=j,X_t^D,\theta] \]

\[ \beta_t(j) = \alpha_t(j) \prod_{i=1}^{S} \beta_{t+1}(i) \times \Pr[s_{t+1}=i|x_{t+1},\theta] \]

The Baum–Welch algorithm estimates weather state occurrence probabilities for each time step, as shown in Equation (A1). Such probabilities can be used to obtain averaged atmospheric patterns corresponding to the conditional precipitation probability distributions, thus allowing a physical interpretation of the model-defined weather states.

**APPENDIX B. THE VITERBI ALGORITHM**

Finding the most likely hidden weather state sequence \( S_t^D \), given the time series \( R_t^D \) and \( X_t^D \), and the model parameters \( \theta \), means solving the maximization problem:

\[ S_t^D = \text{arg max}_{S_t^D} \Pr[S_t^D|R_t^D,X_t^D,\theta] \]

which is equivalent to maximizing \( \Pr[S_t^D,R_t^D|X_t^D,\theta] \). Let us define the variable \( \omega_t(j) \), which is the highest probability of ending in state \( j \) at time \( t \), while having observed \( R_t^D \) conditionally on \( X_t^D \) and \( \theta \). The probabilities \( \omega_t(j) \) (for \( t = 1,\ldots,D \) and \( j = 1,\ldots,S \)) are obtained by maximizing over \( S_t^{D-1} \):

\[ \omega_t(j) = \max_{S_t^{D-1}} \Pr[s_t=j,S_t^{D-1},R_t^D|X_t^D,\theta] \]

which is computed by induction:

\[ \omega_t(j) = \max_i \omega_{t-1}(i)\Pr[s_t=j|s_{t-1}=i,x_t,\theta]f_t^T(r_t|s_t=j,E_t,\theta) \]

if \( t > 1 \)

Moreover we need to define the tracking variable \( \psi_t(j) \) (for \( t = 1,\ldots,D \) and \( j = 1,\ldots,S \)), which is the most likely state to occur at time \( t-1 \), given that state \( j \) occurs at time \( t \). Thus \( \psi_t(j) \) is equal to the argument that maximizes the right-hand side of Equation (B3):

\[ \psi_t(j) = \text{arg max}_{i} \omega_{t-1}(i)\Pr[s_t=j|s_{t-1}=i,x_t,\theta] \]

if \( t > 1 \)

To carry out the complete calculations of the Viterbi algorithm, \( \omega_t(j) \) and \( \psi_t(j) \) have to be initialized:

\[ \omega_1(j) = \Pr[s_1=j|x_1,\theta]f_1^T(r_1|s_1=j,E_1,\theta) \]

\[ \psi_1(j) = 0 \]

After initialization, \( \omega_t(j) \) and \( \psi_t(j) \) are computed recursively via Equations (B3) and (B4). Finally the most likely state sequence is determined by path backtracking. First, the most likely state at the last time step \( D \) is determined as the argument that maximizes \( \omega_D(j) \):

\[ \hat{s}_D = \text{arg max}_{j} \omega_D(j) \]

Then \( \hat{s}_t^{D-1} \), which is the remainder of the most likely state sequence, is determined recursively for \( t = D-1,\ldots,1 \):

\[ \hat{s}_t = \psi_{t+1}(\hat{s}_{t+1}) \]