

# Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing

M. Vrac<sup>1,5,\*</sup>, M. Stein<sup>2</sup>, K. Hayhoe<sup>3,4</sup>

<sup>1</sup>Center for Integrating Statistical and Environmental Science, The University of Chicago, 5734 S. Ellis Avenue, Chicago, Illinois 60637, USA

<sup>2</sup>Department of Statistics, The University of Chicago, 5734 S. University Avenue, Chicago, Illinois 60637, USA

<sup>3</sup>Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, 105 S. Gregory Street, Urbana, Illinois 61801, USA

<sup>4</sup>Department of Geosciences, Texas Tech University, Lubbock, Texas, USA

<sup>5</sup>*Present address:* Laboratoire des Sciences du Climat et de L'Environnement, Centre d'Etudes de Saclay, Orme des Merisiers, 91191 Gif-sur-Yvette, France

**ABSTRACT:** We present a novel statistical downscaling method that provides accurate and relatively transparent simulations of local-scale precipitation characteristics. The method combines large-scale upper-air circulation with surface precipitation fields, and is based on a nonhomogeneous stochastic weather typing approach. Here we apply the method to downscale precipitation at 37 rain gauges in the state of Illinois, USA. Regional climate conditions are categorized in terms of 2 different types of weather states: (1) 'precipitation patterns' developed by a hierarchical ascending clustering (HAC) method with an original metric applied directly to the observed rainfall characteristics in Illinois, and (2) 'circulation patterns' developed by a mixture model applied to large-scale NCEP reanalysis fields. We modeled the transition probabilities from one pattern to another by a nonhomogeneous Markov model that is influenced by large-scale atmospheric variables such as geopotential height, humidity and dew point temperature depression. Our results indicate that including the precipitation states in the statistical model allows us to simulate important precipitation features such as conditional distributions of local simulated rainfall intensities and wet/dry spell behavior more accurately than with a traditional approach based on upper-air circulation patterns alone.

**KEY WORDS:** Weather states · Weather typing · Nonhomogeneous Markov model · Hierarchical clustering · EM algorithm · Statistical downscaling

—Resale or republication not permitted without written consent of the publisher—

## 1. INTRODUCTION

### 1.1. Downscaling climate change projections

Translating global model simulations of future climate change into high-resolution regional-scale projections represents one of the primary challenges to climate change impact assessments. Barriers to generating these high-resolution projections are computational in nature, as well as being due to limitations in our understanding of the geophysical processes that determine the interactions between larger-scale patterns of change and local climate. Resolving future changes in surface climate characteristics at the local

level, however, is essential to determining the likely impacts of climate change. The value of climate information to regional planners and decision makers increases significantly and may even be altered by the spatial resolution of the analysis (e.g. Mearns 2003 and references therein). The more precise (and hopefully more accurate) our ability to determine what changes will occur and where, the more effective the allocation of resources to address potential impacts from those changes. Hence, extraction of higher-resolution spatial climate information from global or regional models has important applications to both long-term forecasting (from a few months to several years) as well as climate impact assessments (on time scales of years to decades).

\*Email: mathieu.vrac@cea.fr

Introduction of spatial variability into larger-scale fields is commonly achieved by either (1) dynamical downscaling, also known as regional climate modeling, or (2) statistical downscaling, which relies on observed relationships between large- and small-scale climate fields to extrapolate into the future (e.g. Mearns et al. 1999, Murphy 1999).

Dynamical downscaling is based on the simulation of the smaller-scale dynamical processes that control climate at the regional level. Coupled atmosphere–ocean general circulation model (AOGCM) output is used to force the boundary conditions of regional models, which can then produce climate projections down to resolutions of 5 km<sup>2</sup>. However, regional climate modeling requires a significant investment in production of both high temporal resolution AOGCM output, as well as the computing resources required to generate the regional model simulations themselves. As such, it is currently possible to apply regional models only to limited periods and regions, and only for the AOGCMs and emission scenarios from which the required high-resolution input fields are available.

Statistical downscaling is the second approach commonly taken to introduce higher spatial resolution into large-scale fields. Relationships between local-scale surface conditions and large-scale variable output fields (chosen such that they are strongly related to the local-scale conditions of interest) are first developed, based on observed data. These relationships are then used to estimate projected changes at higher spatial resolutions (at the station level or at a finer grid-scale level), based on future projections of larger-scale fields by global or regional models.

### 1.2. Weather typing and clustering methods for downscaling

The weather typing approach to statistical downscaling generally attempts to define weather states based on large-scale upper-air variables alone, and has demonstrated some success in reproducing observed surface precipitation patterns, particularly when combined with stochastic weather generation (e.g. Wilby et al. 1998, 2002, Wilks 1999, Wilks & Wilby 1999, Bischoff & Vargas 2003). Atmospheric patterns are generally determined in 2 ways, either through subjective or objective definitions. The subjective approach to defining weather types is based on prior knowledge and expert opinion, the best known being the Lamb Weather Type (LWT) classification for the British Isles (Lamb 1972). In contrast, the objective approach determines classifications for a certain region automatically on the basis of clustering methodologies.

Many clustering and classification methods have been applied in atmospheric sciences in general and in weather typing in particular. For instance, Zorita et al. (1993) employed a hierarchical descending clustering method applied to the first few components of an empirical orthogonal function (EOF) analysis of sea level pressure anomalies to identify 3 weather states in the Columbia River Basin for the winter season. Schnur & Lettenmaier (1998) used a similar approach for sea level pressure over Australia to define stochastic models of rainfall amounts according to weather states.

A 'K-means' clustering procedure (sometimes termed 'dynamic clustering') was used by Akkur et al. (1992) as part of a framework to predict ozone in Australia and by Bárdossy et al. (1993) for stochastic precipitation modeling. Huth (2001) performed a 'repeated' K-means approach (keeping the recurrent clusters) on daily 500 mb geopotential height fields to define circulation types over Europe, and related these to 9 climate variables in the Czech Republic. In order to derive a climatology of severe storms in Virginia, Davis et al. (1993) used results from an EOF analysis of surface and upper-air observations, such as sea level pressure, wind speed and direction, temperature and dew point at the surface and at 850, 700 and 500 mb constant pressure surfaces, to obtain clusters via a hierarchical agglomerative algorithm. These clusters were used as the initial step in a K-means clustering procedure, which produced 9 circulation patterns or clusters. Similar methods have been applied to surface variables, for instance by Kalkstein & Corrigan (1986), to derive a synoptic climatology for sulfate concentrations in Delaware, and by Davis & Kalkstein (1990) to define air quality regimes.

Neural networks, generally employed as universal regression tools, can be used in a clustering context through self-organizing maps (SOMs), e.g. Hewitson & Crane (2002), to downscale daily precipitation. They can also be used for classification purposes; e.g. Bárdossy et al. (1993) used 51 input neurons, 2 hidden layers of 45 and 40 neurons and 20 output neurons to classify 700 mb pressure surfaces according to a given scheme. Diday & Vrac (2005) developed a statistical mixture of 'distributions of distributions' through the copulas theory to determine air mass types from vertical atmospheric profiles of specific humidity and temperature (Naveau et al. 2004, Vrac et al. 2005). Pongracz et al. (2001) used a fuzzy rule approach to model frequency distributions of monthly precipitation in Hungary, conditional on the Hess-Brezowsky circulation patterns. Because large-scale seasonal cycles have obvious influences on local-scale climate features, Vrac et al. (2006) also applied an 'estimation-maximization' (EM) technique to identify and compare seasonal atmospheric patterns and their temporal shifts over North America from reanalysis and AOGCM outputs.

### 1.3. Our approach

Downscaling methods based on patterns in upper-air variables help to understand climate at the global scale (i.e. at the pattern scale). However, defining the statistical relationships between the states and some local surface variables can be complicated, and local variables of interest are rarely considered in the definition of the weather states. Rather, the relationships between local variables and large-scale patterns are developed after the patterns have been identified.

Here, we address the issue of global linkages to local-scale precipitation features through application of a stochastic approach based on a nonhomogeneous hidden Markov model. This approach, though inspired by the initial work of Hughes & Guttorp (1994) and Hughes et al. (1999), includes the following significant modifications: (1) We placed the method in a weather typing scheme by defining prior weather states or circulation patterns, but retaining the stochastic modeling of the precipitation. Hence, we characterized not only the states themselves, but also the probabilities of transition between those states, using a nonhomogeneous Markov model (NMM) that is not hidden anymore. (2) We selected 2 different types of weather states and used these to evaluate the stochastic weather typing approach—‘classical’ upper-air patterns based on large-scale atmospheric variables, and original precipitation-related patterns based on available observed local rainfall intensities in the studied region; the latter patterns correspond to weather states that are more regional- than large-scale, and were directly defined from the climate variable to be downscaled. After developing this approach, we evaluated the performance of large-scale circulation pattern-based downscaling vs. the new downscaling approach that combines large-scale circulation fields (used to determine the transition probabilities between patterns) with regional precipitation-based patterns. Based on NCEP reanalysis output fields, we downscaled precipitation to 37 long-term rain-gauges in the state of Illinois, USA, and compared them with observed station-based precipitation records using a 20 yr dataset independent of the 20 yr dataset used to fit the downscaling model, in order to draw conclusions regarding the relative performance of the different downscaling methods.

## 2. STATISTICAL MODEL AND METHODS

Hughes & Guttorp (1994) first suggested the use of a nonhomogeneous hidden Markov model (NHMM) to characterize the transition between non-observable weather states. Conditional on the current state and on

each site, Hughes et al. (1999) modeled precipitation occurrences through an autologistic model for multivariate binary data for a network of rain gauges. Likewise, given the current state and a site, Bellone et al. (2000) applied a mixture of a Gamma distribution and a point mass at zero (both site- and state-specific) to describe the distribution of the rainfall intensity. Although the latter assumes spatially independent structures—i.e. except for the weather state that is common to each location, nothing relates one site to another—the results appear realistic when applied to a test set of observational data. The main difference to a more ‘classical’ weather typing method is that, in this approach, the states are ‘precipitation’ patterns and are obtained *a posteriori* rather than prior to modeling conditional distributions. As a consequence, the patterns provide a good description of local precipitation structure. However, if we consider large-scale variables such as geopotential height or wind, the statistical distributions of these variables can overlap, reducing the interpretation of these weather states in terms of large-scale conditions. In contrast, a classical weather typing method might not describe precipitation as successfully, but the global circulation patterns would be more distinct.

Inspired by the work of Hughes & Guttorp (1994) and Bellone et al. (2000), we developed a conceptually different alternative to NHMM. Our main goal was to include their stochastic modeling of precipitation—with the main assumption of a NMM—in a more classical weather typing framework, i.e. by defining ‘prior’ weather states or weather patterns, not determined through a hidden Markov model. This context makes this novel method arguably easier to interpret because of the lack of a hidden layer. We then compared 2 approaches, one where our new method is conditioned on regional precipitation patterns, and another where it is conditioned on large-scale circulation patterns.

### 2.1. Nonhomogeneous stochastic weather typing model (NSWT)

In describing our model, we use the following naming conventions. Let  $\mathbf{R}_t = \{R_t^1, \dots, R_t^n\}$  be a multivariate random vector giving rainfall intensities at a network of  $n$  weather stations at daily Time  $t$ ,  $1 \leq t \leq T$ . Let  $S_t$  be the weather state at Time  $t$ , and  $S_1^{t-1}$  be the sequence of states from Time 1 to  $t-1$ . Let  $\mathbf{X}_t$  be a vector of atmospheric variables at Time  $t$ , and  $\mathbf{X}_1^T$  the sequence of values of the atmospheric variables of  $\mathbf{X}_t$  from Time 1 to Time  $T$ .

Our nonhomogeneous stochastic weather typing model is defined by 2 assumptions. Assumption (1) can be formulated as in Hughes et al. (1999) by

$$P(S_t | S_{t-1}^{-1}, \mathbf{X}_t^\top) = P(S_t | S_{t-1}, \mathbf{X}_t) \quad (1)$$

Eq. (1) means that the probability of being in a given state at Time  $t$  depends only on the state at previous Time  $t-1$  and on the value of the vector of large-scale atmospheric variables at current Time  $t$ . A simple way to parameterize this probability is as follows:

$$P(S_t = j | S_{t-1} = i, \mathbf{X}_t) \propto \gamma_{ij} \exp[-\frac{1}{2} (\mathbf{X}_t - \mu_{ij}) \Sigma^{-1} (\mathbf{X}_t - \mu_{ij})'] \quad (2)$$

where  $\Sigma$  is the variance-covariance matrix for the atmospheric data (centered around their mean),  $\mu_{ij}$  is the mean vector of the atmospheric variables when the weather state at the previous time was  $i$  and the current state is  $j$ , and  $\gamma_{ij}$  gives the baseline transition probability from State  $i$  to State  $j$ .

It is clear that an NMM is a generalization of a homogeneous Markov model. Indeed, the exponential term in Eq. (2) corresponds to the influence of the atmospheric variables  $\mathbf{X}_t$  in the state transition probabilities. If we remove this term, we retrieve a homogeneous Markov model. So this term is used to influence the appearance of the states (for either historical simulations or future projections) based on upper-air variables, i.e. not just from the given transition probabilities of a Markov model.

Most of the time, the distribution of the large-scale variables  $\mathbf{X}_t$  we use is symmetrical. Hence, for convenience, a Gaussian distribution is employed in Eq. (2), even though other models could have been used. Eq. (1) assumes that the transition probabilities of the Markov model are proportional to the conditional density of  $\mathbf{X}_t$  given the transition. This assumes that the transition probabilities and therefore the appearance of the weather states will change under conditions of global change affecting the atmospheric variables  $\mathbf{X}_t$ .

Assumption (2) in our NSWT approach can be written as

$$f_{\mathbf{R}_t | S_t^\top, \mathbf{R}_1^{t-1}, \mathbf{X}_t^\top}(r) = f_{\mathbf{R}_t | S_t, \mathbf{X}_t}(r) \quad (3)$$

where  $f$  denotes a probability density function and  $\mathbf{R}_1^{t-1}$  is the sequence of precipitation amounts from Time 1 to  $t-1$ . Eq. (3) says that, given the weather state, the observed stochastic process  $\mathbf{R}_t$  (rainfall) is assumed to be conditionally temporally independent. The parameterization can be done by modeling precipitation at each station, given the weather state, as a mixture of a point mass at zero and a gamma distribution (with site-specific parameters) as follows:

$$f_{\mathbf{R}_t | S_t, \mathbf{X}_t}(r_t) = \prod_{i=1}^n [p_{si}(\mathbf{X}_t) G(r_t^i | \alpha_{si}, \beta_{si})]^{1_{\{r_t^i > 0\}}} [1 - p_{si}(\mathbf{X}_t)]^{1_{\{r_t^i = 0\}}} \quad (4)$$

where  $n$  is the number of rain stations,  $p_{si}$  is the precipitation probability at Stn  $i$  in State  $s$ , and  $r_t^i$  is the precipitation amount at Stn  $i$  and Time  $t$ . The indicator function  $1_{\{r_t^i > 0\}}$  takes the value 1 if the precipitation value at Time  $t$  and Stn  $i$  is  $>0$  and takes the value 0

if  $r_t^i = 0$ . The density of the gamma distribution at  $r_t^i$  is denoted by  $G(r_t^i | \alpha_{si}, \beta_{si})$  with parameters  $\alpha_{si}$  and  $\beta_{si}$  depending on State  $s$  and Stn  $i$ :

$$G(r | \alpha_{si}, \beta_{si}) = \frac{\beta_{si}^{\alpha_{si}}}{\Gamma(\alpha_{si})} r^{\alpha_{si}-1} \exp(-r\beta_{si}) \quad (5)$$

where  $\Gamma(\cdot)$  is the gamma function.

This model means that, for each weather station, we model as many distributions of rainfall intensity and probabilities of rain occurrence as the number of weather states. The product in Eq. (4) corresponds to a spatial independence of the rainfall occurrences and intensities, conditional on the weather state. Although the NHMM approach developed by Bellone et al. (2000) used the same first assumption defined by Eq. (1), it did not condition the density of rainfall intensity  $\mathbf{R}_t$  on  $\mathbf{X}_t$  as we do here in our second assumption (Eq. 3). Hence, given State  $s$  and Stn  $i$ , Bellone et al. (2000) used constant probabilities  $p_{si}$  of precipitation. In practice, due to the variations of rain rate caused by the possible presence of different precipitation regimes, these constants can be unrealistic (Jeffries & Pfeiffer 2000). That is why, in Eq. (4), we model these probabilities as functions of large-scale atmospheric variables  $\mathbf{Y}_t$ , through logistic regression models:

$$p_{si}(\mathbf{Y}_t | \lambda_{si}) = P(1_{\{r_t^i > 0\}} = 1 | S_t = s, \mathbf{Y}_t) = \frac{\exp(\mathbf{Y}_t' \lambda_{si})}{1 + \exp(\mathbf{Y}_t' \lambda_{si})} \quad (6)$$

where  $\lambda_{si}$  (which is state- and site-specific) is the vector of unknown logistic regression coefficients. The vector  $\mathbf{Y}_t$  may or may not have components in common with the atmospheric variables  $\mathbf{X}_t$  employed to influence the transition probabilities in Eq. (1). For convenience, we will assume  $\mathbf{Y}_t = \mathbf{X}_t$  in this analysis and associated applications. Hence, the estimation process provides parameters for the rainfall occurrence model (Eq. 6), not constant numerical values of probabilities.

As for the transition probabilities, conditioning the rainfall occurrences on  $\mathbf{X}_t$  implies that the occurrence probabilities will be altered under conditions of global change affecting the atmospheric variables  $\mathbf{X}_t$ . If  $\mathbf{X}_t$  is correctly chosen, the rainfall occurrence probabilities should be adequately driven and estimated. The rainfall intensity itself is not regressed directly against atmospheric variables (i.e. the parameters and the associated model itself do not depend on large-scale atmospheric variables—except on the weather states—as do the occurrences). This arguable choice, made for simplicity of the model, implies that to characterize the influence of large-scale atmospheric features on local rainfall intensity, it is sufficient to use gamma distributions with state-specific parameters, with weather states that summarize recurrent large-scale patterns.

## 2.2. Fitting the NSWT model to the weather patterns

By specifying the weather states (i.e. patterns) *a priori*, the state 'labels' are known and the model can be trained by straight maximum likelihood, without requiring the EM algorithm (Dempster et al. 1977). More precisely, once the patterns are achieved *a priori*, a NMM—Eqs. (1) & (2)—is fitted to the patterns to represent the transition process from one daily pattern to another. The NMM (that is not hidden) is applied to characterize the observed temporal sequence of the patterns, rather than to define them as in the NHMM approach. Since the state labels are known, the parameters in Eqs. (4) to (6) are estimated for the fitting period (1980–1999) conditionally on the obtained weather states, through a direct maximization of the likelihood for the model defined by Eq. (4).

The set of atmospheric predictors in this study was selected based on previous studies (e.g. Charles et al. 1999) which found that local precipitation is closely connected to specific large-scale atmospheric fields, as well as on the practical issue of which daily fields were available from reanalysis for use in this study. The vector of large-scale atmospheric variables  $\mathbf{X}_t$  is defined here in terms of 3 NCEP reanalysis variables at 850 mb: geopotential height  $Zg_{850}$ , specific humidity  $Q_{850}$ , and dew point temperature depression  $\Delta Td_{850}$  (defined as the difference between the temperature  $T_{850}$  and the dew point temperature  $Td_{850}$ ):

$$\Delta Td_{850} = T_{850} - Td_{850} \quad (7)$$

In particular, dew point temperature depression is an index of the moisture saturation of the atmosphere and is very relevant for precipitation downscaling (Charles et al. 1999).  $Zg_{850}$  was introduced in our model, because this variable is frequently used in defining scale relationships for downscaling purposes. After examining other variables,  $Q_{850}$  was also included, because it noticeably increased the quality of the downscaled precipitation estimates, compared to observations.

$\mathbf{X}_t$  was not directly defined in terms of these 3 upper-air variables. Instead, for any one of the atmospheric variables, we defined  $C$  as the matrix with element  $c_{ij}$  representing the correlation coefficient between the rainfall intensity time series at Stn  $i$  and some atmospheric variable, e.g.  $Zg_{850}$ , at grid-cell  $j$ . To relate the occurrence of rain at Stn  $i$  in Illinois to large-scale information, it would be meaningless to consider the direct correlation between rainfall at Stn  $i$  and a very large number of NCEP grid-cells covering a region comprising Illinois. Instead, we assumed that the weather patterns (precipitation or circulation patterns; see Sections 2.3 & 2.4) influence the atmospheric variables over Illinois, which, *à leur tour*, influence rainfall occurrence at Illinois weather stations. Even if it is reasonable to

assume that, in case of rain at Stn  $i$ , the intensity of local rainfall—e.g. heavy or light rainfall, as characterized by the parameters of the gamma distribution—depends largely on the current large- or regional-scale weather pattern, it is also reasonable to assume that the event 'rain' at Stn  $i$  is driven only by atmospheric variables over Illinois. That is why we only considered  $m = 6$  grid cells covering Illinois for the correlation matrix. For each of the 3 NCEP variables selected, the associated  $N \times m$  correlation matrix  $C$  was then decomposed by applying a singular value decomposition (SVD) step after von Storch & Zwiers (1999) as follows:

$$C = UDV' \quad (8)$$

where  $U$  is an  $N \times N$  matrix,  $V$  an  $m \times m$  matrix,  $V'$  the transpose of  $V$ , and  $D$  is an  $N \times m$  diagonal matrix whose diagonal elements ( $d_1, \dots, d_m$ ) are the singular values of  $C$ . A summary of the original atmospheric variable can be built by multiplying the atmospheric variable, standardized for each grid cell separately, by Column  $i$  of Matrix  $V$ . This new summary variable contains  $(d_i^2 / \sum_{j=1}^m d_j^2) \times 100\%$  of the correlation between the atmospheric variable and the rainfall process.

We applied this SVD step separately to each of the 3 NCEP variables  $Zg_{850}$ ,  $Q_{850}$  and  $\Delta Td_{850}$ . For each variable, we retained the first summary, representing 93.6, 98.6 and 97.5% of the correlation, respectively. This results in 3 summarizing variables instead of the initial 18. The vector  $\mathbf{X}_t$ , used in Eqs. (2) & (6), contains these 3 summaries.

Our goal was not only to develop a transparent pattern-based downscaling approach, but also to compare 2 approaches where our model is conditioned on the patterns of (1) regional precipitation, and (2) large-scale circulation. For this reason, before examining the results obtained by applying the method to large-scale NCEP reanalysis fields, we next describe the methods used to define the precipitation and circulation patterns.

## 2.3. Defining precipitation patterns

When downscaling daily surface precipitation, it is logical to assume that patterns directly based on measured precipitation will prove more accurate than those that are indirectly derived. For this reason, we first defined a downscaling method based solely on precipitation patterns that describe surface rainfall for the study region, without considering upper-air variables. That is, we used rainfall measured at  $N$  weather stations (rather than large-scale NCEP reanalysis fields) to determine precipitation patterns, by clustering the  $T$  days of the fitting time period (see Section 3 for data used).

Here, a hierarchical ascending clustering method (HAC) was used. This approach does not assume any

(explicit) distribution for the variables. Starting with the items to be clustered as groups with only 1 element (singletons), the principle consists, at each step, of regrouping 2 groups according to a given criterion, until we only have 1 group, comprising all the items and using the Ward criterion (Ward 1963) to maximize the interclass variance.

The main problem with using the HAC approach, however, is that if we only use raw (measured) values of rainfall to characterize each day, the HAC method (and almost any clustering method) will only consider of the 'high' values of rainfall, or at least those that are not too close to zero. Moreover, to obtain patterns that differ in more than their mean rainfall intensity, we need to introduce spatial and temporal precipitation characteristics into the clustering process.

For these reasons, we applied the HAC method using an original metric  $D$  to evaluate the degree to which 2 days  $t$  and  $t'$  differ, instead of the more classical Euclidian distance. The metric  $D$  is defined as

$$D(P_t, P_{t'}) = \sum_{i=1}^N d(P_{ti}, P_{t'i}) \quad (9)$$

where  $P_t = (P_{t1}, \dots, P_{tN})$  corresponds to the precipitation characteristics for each of the  $N$  weather stations. The characteristics of Stn  $i$  for Day  $t$  are  $P_{ti} = (R_{ti}, Sa_{ti}, Ta_{ti})$ , where  $R_{ti}$  is the measured rainfall intensity, and  $Sa_{ti}$  and  $Ta_{ti}$  are the spatial and temporal rainfall anomalies, respectively, defined by  $Sa_{ti} = R_{ti} - 1/N \sum_{i=1}^N R_{ti}$  and  $Ta_{ti} = R_{ti} - 1/T \sum_{t=1}^T R_{ti}$ . Thus,  $Sa_{ti}$  gives the spatial variation on Day  $t$  and  $Ta_{ti}$  gives the temporal variation for Stn  $i$ . The metric  $d$  in Eq. (9), between 2 sets of characteristics  $P_{ti}$  and  $P_{t'i}$  for 2 days  $t$  and  $t'$  for a given Stn  $i$ , is given by

$$d(P_{ti}, P_{t'i}) = \begin{cases} 0 & \text{if } R_{ti} = R_{t'i} = 0 \\ E_d[(Sa_{ti}, Ta_{ti}), (Sa_{t'i}, Ta_{t'i})] & \\ + h(R_{ti}, R_{t'i}) + \alpha \mathbb{1}_{\{R_{ti}=0\} \neq \mathbb{1}_{\{R_{t'i}=0\}}} & \text{otherwise} \end{cases} \quad (10)$$

where  $E_d$  is the bivariate Euclidean distance and  $h$  is the function

$$h(x, y) = |\log(x + \epsilon \mathbb{1}_{\{x=0\}}) - \log(y + \epsilon \mathbb{1}_{\{y=0\}})| \quad (11)$$

where  $|\cdot|$  corresponds to the absolute value and  $\epsilon = 10^{-3}$ . The value of  $10^{-3}$  is somewhat arbitrary and was chosen because the difference between  $\log(m)$  ( $m$  = minimum observable rainfall) and  $\log(10^{-3})$  is about 3-fold larger than the difference between  $\log(m)$  and  $\log(m_2)$  ( $m_2$  = second minimum observable rainfall);  $m$  is described in Section 3).

With the metric given in Eqs. (9) & (10), we consider that the distance between the characteristics  $P_{ti}$  and  $P_{t'i}$  is zero if and only if the 2 raw values of rainfall are zero. To define precipitation patterns, we do not want to distinguish between 2 null values of rainfall. If they are not both equal to zero, we compute a positive metric value by adding 3 terms: (1)  $E_d[(Sa_{ti}, Ta_{ti}), (Sa_{t'i}, Ta_{t'i})]$ , the bi-

variate Euclidean distance applied to the pairs of spatial and temporal anomalies; this allows the clustering algorithm to define precipitation structures that are influenced by the spatial and temporal information at our disposal. (2)  $h(R_{ti}, R_{t'i})$ , the absolute value of a log-difference directly applied to the raw value of precipitation; the introduction of the log function allows us to emphasize small values of precipitation. (3)  $\alpha$  is a penalty term added when only 1 of the 2 raw precipitation values is zero; this produces an improved distinction between positive and zero values of rainfall. It is important to make a good distinction between small values, but essential in the case where 1 value is zero. The value of 1 was chosen for  $\alpha$ ; based on the data (Section 3), it appears to provide a reasonable penalty compared to the mean value given by  $E_d$ , producing the clearest and most meaningful patterns.

It is also important to note that metric  $d$  (Eq. 10) is not scale invariant. Here, it is defined with units in cm. If other units are used, this metric has to be redefined. For example,  $E_d$  and  $h$  do not give equivalent results in cm as compared to mm. Consequently, if the units in the analysis are changed, the value of  $\alpha$  must be adjusted to provide a reasonable penalty.

## 2.4. Defining circulation patterns

The method used to define our circulation-based weather states was a mixture of statistical distributions solved by an EM-type algorithm (Dempster et al. 1977, review by McLachlan & Peel 2000). Although it is considered a very efficient clustering method, few studies have used it to define atmospheric patterns (e.g. Smyth et al. 1999). Recently, Vrac et al. (2007) showed that the statistical mixture clustering method tends to produce more consistent results than a hierarchical clustering approach across various datasets (e.g. NCEP vs. ERA-40 reanalysis), and it is more sensitive to day-to-day variations in pattern frequencies, even within seasons such as winter and summer that are characterized by a strong signal.

Here, we assumed that the distribution of the large-scale atmospheric variables under consideration is a statistical mixture of several components, each one characterizing the distribution associated with a single circulation pattern, i.e. one component of the mixture. That is, if  $g$  is the probability density function (PDF) of a large-scale atmospheric variable such as geopotential height at 850 mb, we can write  $g$  as a weighted sum of  $K$  parametric PDFs:

$$g(x) = \sum_{i=1}^K \pi_i g_i(x; \alpha_i) \quad (12)$$

where  $g_i$  is a parametric PDF with parameter  $\alpha_i$  associated with component  $i$  of the mixture,  $\pi_i$  is the 'mixture

ratio' and corresponds to the probability of belonging to component  $i$ , and  $K$  is the number of components (here, circulation patterns). To solve Eq. (12) we use an EM algorithm consisting of 2 successive and iterative steps of Expectation (E step) and Maximization (M step) of the so-called complete log-likelihood. For convenience, the parametric PDFs  $g_i$  are assumed to be Gaussian distributions, and  $\alpha_i = (\mu_i, \Sigma_i)$  where  $\mu_i$  is a vector of means and  $\Sigma_i$  the variance-covariance matrix of  $g_i$ . Note that  $\mu_i$  and  $\Sigma_i$  are different from those in Eq. (2). A constraint is imposed on the variances in the maximization process (M step) to avoid singularities of the likelihood. These singularities occur when a vector of means of 1 component, e.g.  $\mu_i$ , is set equal to any observed multidimensional data, and the variances tend to zero.

Because circulation patterns usually involve a large number of grid cells, we need to reduce the dimension of the problem. This was done by applying a principal component analysis (PCA) to the dataset to be clustered. Before the PCA, the data were standardized so that each statistical variable had a mean equal to 0 and variance equal to 1. Principal components representing 99% of the variance were retained to run the EM algorithm. The number of components  $K$  is usually given *a priori*, but in our application (Section 3), the Bayesian Information Criterion (BIC) served to determine the 'best'  $K$ . We tested values from  $K = 1$  to  $K = 10$ . Each EM run attempts to fit the model in Eq. (12) to the data not only for a given value of  $K$ , but also for several given structures of the variance matrix of the  $K$  components. For instance, the variance matrix can be spherical, diagonal or ellipsoidal, and with equal or varying volumes (see Fraley & Raftery 2002). The BIC was calculated for each pair of a variance structure and a value  $K$ , and the pair (variance,  $K$ ) minimizing the BIC was retained. Hence,  $K$  was obtained as a compromise between a high likelihood with a very complex model (with too many parameters) and a simple model with a low likelihood. From the estimated parameters, patterns were then derived by applying the principle of posterior maximum, with

$$S_i = \{x : \pi_i g_i(x, \alpha_i) \geq \pi_k g_k(x, \alpha_k), k = 1, \dots, K\} \quad (13)$$

where  $S_i$  denotes the circulation pattern corresponding to State  $i$ .

### 3. APPLICATION OF NONHOMOGENEOUS STOCHASTIC WEATHER TYPING (NSWT) TO PRECIPITATION DOWNSCALING

#### 3.1. Station and reanalysis data

Two distinct 20 yr time periods were used to fit and evaluate the statistical method (Eq. 4) designed to

introduce fine-scale spatial variability into larger-scale fields. The fitting period was January 1, 1980 to December 31, 1999, and the period used to evaluate the downscaled results was January 1, 1952 to December 31, 1971. For each of these periods, 2 primary datasets were used. The first consisted of daily precipitation measurements at 37 weather stations in Illinois, and the second of upper-air reanalysis data ( $2.5 \times 2.5^\circ$  spatial resolution) from the National Center for Environmental Prediction (NCEP). Variables used to define large-scale patterns included geopotential height ( $Zg$ ), temperature ( $T$ ), specific humidity ( $Q$ ), relative humidity ( $H$ ) and dew point temperature ( $Td$ ); see Section 2.2 for definition of variables. We focused on the winter season (December, January, February), which represents 1805 d for each of the two 20 yr periods.

#### 3.2. Evaluation of NSWT based on precipitation patterns

We first evaluated the ability of the NSWT method to downscale daily winter precipitation based on patterns that describe surface rainfall only. Upper-air variables are still used to determine the transition probabilities between weather states (see Section 2.2). However, we used station-based observed rainfall at each of the 37 Illinois stations (rather than large-scale NCEP reanalysis) to determine regional precipitation patterns, by clustering the days from the 20 winters of the fitting time period (1980–1999).

The metric given in Eq. (10) was employed in the HAC method using the Ward criterion to choose the number of clusters, with the criterion being the stabilization of the interclass variance reduction. According to this criterion, the best partition contained 4 clusters (Fig. 1). Patterns 1 and 4 correspond to the weakest and strongest rainfall intensities, respectively, whereas Patterns 2 and 3 represent moderate precipitation. Pattern 2 shows a southwest–northeast gradient. A north–south gradient is discernible in Patterns 3 and 4, the typical distribution of winter rainfall in Illinois. Visual comparisons with daily maps of precipitation, e.g. as provided by the Illinois State Water Survey (available at: [www.sws.uiuc.edu/atmos/statecli/mapsv2/mapsv2.htm](http://www.sws.uiuc.edu/atmos/statecli/mapsv2/mapsv2.htm)), confirm these spatial gradients, suggesting that the 4 precipitation patterns in Fig. 1 capture the principal features of winter precipitation in Illinois.

Once the statistical parameters have been estimated, local rainfall at each station can be simulated using daily NCEP-based values of large-scale atmospheric summary variables  $\mathbf{X}_t$  ( $Zg_{850}$ ,  $Q_{850}$  and  $\Delta Td_{850}$ , see Section 2.2) to influence transition probabilities from one weather state to another, and to drive precipitation probabilities at each site. Based on the parameters

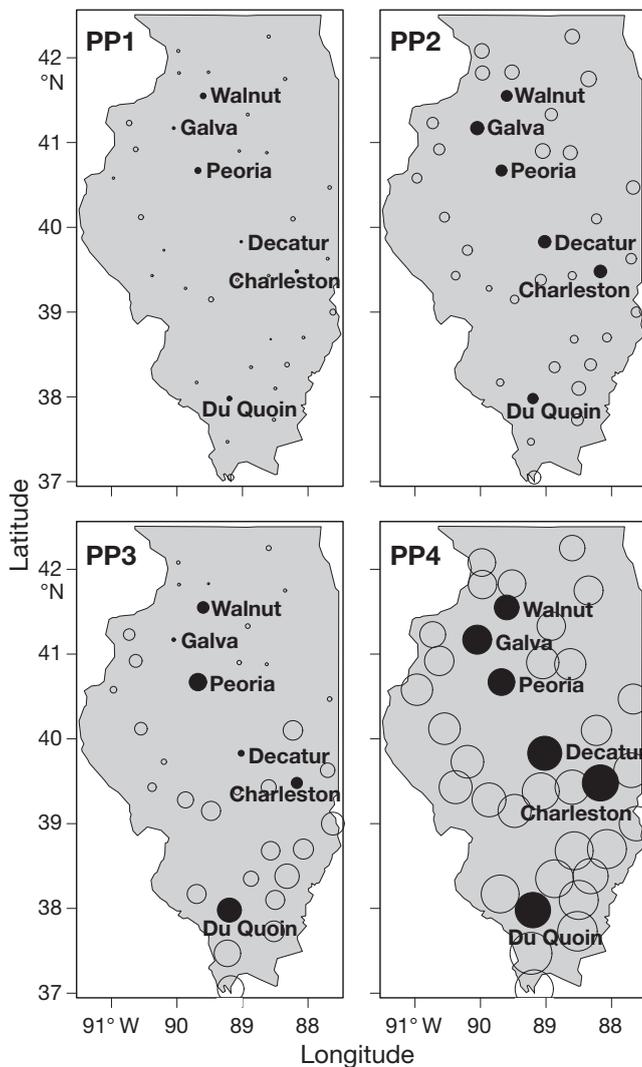


Fig. 1. Station-based precipitation patterns (PP1 to PP4) over Illinois derived by a hierarchical ascending clustering (HAC) method using the Ward criterion combined with the metric  $D$  (Eq. 9) to capture spatial variability among patterns. Area of circle is proportional to mean rainfall. Representative stations are indicated

obtained over the fitting period 1980–1999, we therefore simulated local precipitation for the 37 weather stations for the evaluation period 1952–1971 as a first test of the quality of our model. Only  $\mathbf{X}_t$  values from the evaluation period were used to simulate temporal sequences of precipitation patterns, and then local precipitation. The normalization of the 1952–1971 NCEP data, which is necessary to build the SVD summary values detailed in Section 2.2, was based on the 1980–1999 NCEP means and variances, and not on the evaluation period characteristics (i.e. not on the 1952–1971 means and variances). This prevents the removal of any potential changes in mean and variability from the evaluation period data.

In simulating  $n$  weather states from  $n$  vectors  $\mathbf{X}_t$  of NCEP reanalysis summaries, we first initialized  $S_1$ , the weather state for Day 1, according to e.g. the proportion of each pattern, or the posterior probability of each pattern conditional on  $\mathbf{X}_1$ . Then, for each Day  $t > 1$ , State  $S_t$  was simulated from State  $S_{t-1}$  for Day  $t-1$  and from the vector  $\mathbf{X}_t$  of the NCEP outputs for Day  $t$ . For each Day  $t$  (for which we now had the simulated State  $S_t$ ) we can simulate a local rainfall amount  $R_{ti}$  for each rain-gauge  $i$ . First, we simulated whether or not a rainfall event is present, according to Eq. (6), based on station- and state-specific parameters. If there is no rainfall event, the rainfall amount  $R_{ti}$  is zero. Otherwise,  $R_{ti}$  was simulated according to a Gamma distribution (Eq. 5) with station- and state-specific parameters. Although the 2 simulation processes for weather states and rainfall amounts are presented separately, it is equivalent to perform both at once by simulating the rainfall amounts for each station for Day  $t$  directly after the simulation of State  $S_t$ . Because of the stochasticity of our approach, applying the same sequence of  $\mathbf{X}_t$  values ( $t = 1, \dots, t = T = 1805$ ) twice will produce 2 slightly different sequences of simulated local precipitation. The simulation process was thus performed 500 times. In consequence, we obtained 500 temporal sequences of precipitation patterns for the  $T = 1805$  d. These simulated patterns are then used to generate 500 sequences of local precipitation for the  $N = 37$  weather stations and the 1805 d.

Initial indications of the degree to which the model is successful at reproducing historical precipitation is given by the quantile–quantile plots (QQplots) comparing observed and simulated daily winter precipitation for the period 1952–1971, as shown in Fig. 2 for the 6 representative stations shown in Fig. 1.

Overall, these QQplots show very good agreement between observed and simulated precipitation for most of the 37 stations, particularly below the 99th percentile. Although there are some gaps between observed and simulated quantiles of extreme values of precipitation for some stations (e.g. Du Quoin and Peoria), the simulations are close to the observed distributions, with no significant bias in either direction.

Simulated precipitation (examples in Fig. 3) consistently reproduced the probability of observed wet and dry spells, even for small values of probabilities, corresponding to long periods. Again, while there is a tendency to under-estimate the wet spell probabilities for some sites and over-estimate the dry spell probabilities for others, the shape of the simulated distributions is close to observed data, without significant bias in either direction. Together, QQplots in Fig. 2 and wet/dry spell probabilities in Fig. 3 suggest that the NSWT approach based on precipitation patterns is capable of generating local-scale daily precipitation

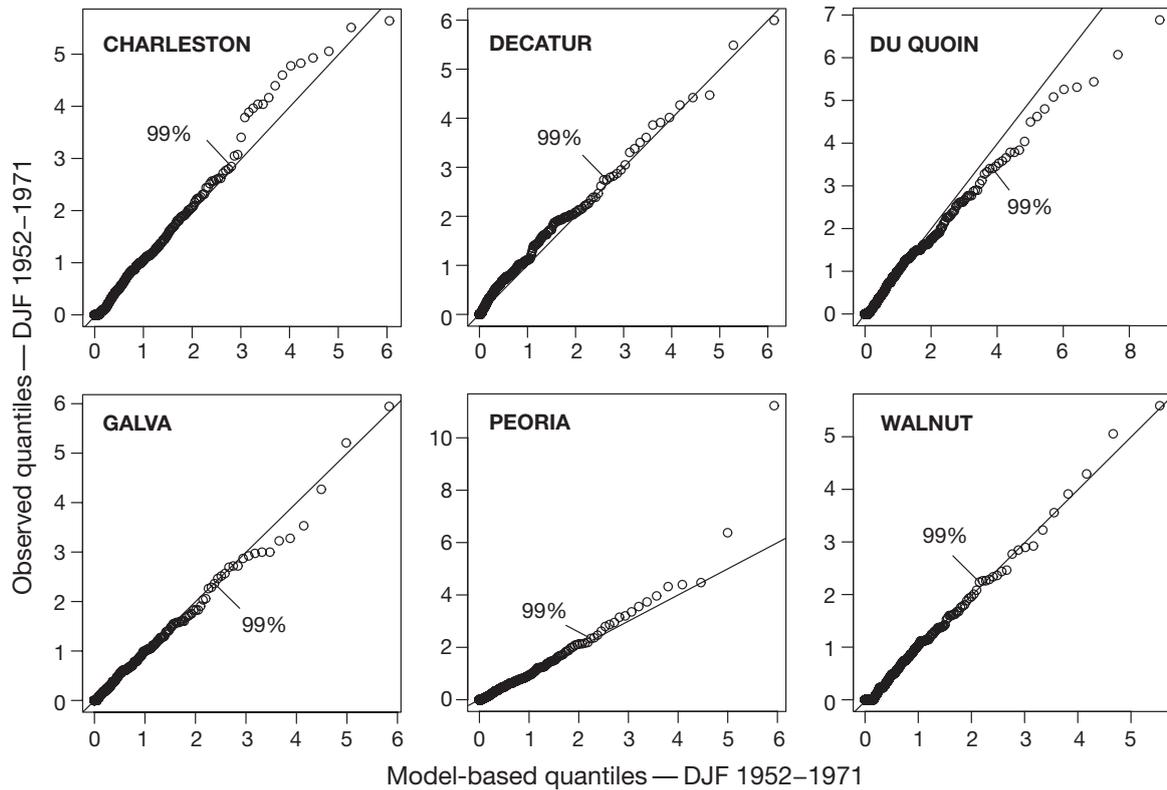


Fig. 2. QQplots of precipitation simulated from 4 precipitation patterns vs. observed precipitation (in cm) for Dec, Jan, Feb 1952–1971; 99th percentiles are indicated

with characteristics similar to observed, even for a period such as 1952–1971, separated by 28 yr from the dataset used for fitting purposes (1980–1999).

To fully evaluate the benefits of using precipitation patterns to drive a NSWT method, we next defined upper-air circulation patterns — normally used in weather typing contexts — and used these to drive the same NSWT downscaling approach.

### 3.3. Evaluation of NSWT based on circulation patterns

We next defined a set of upper-air circulation patterns (Section 2.4). To identify the upper-air fields most relevant to surface winter precipitation over the state of Illinois, we first performed clustering of upper-air reanalysis output fields for a range of different variables; we then interpreted the patterns obtained in terms of large- and/or local-scale variables, and assessed the differences between the patterns within the results from each given set of variables. This assessment was accomplished, for example, through boxplots comparing the relationship of the derived patterns to large- and/or local-scale variables, through modeling/estimation of the relationships and/or corre-

lations to local-scale precipitation, and finally, by simulating local precipitation from relevant sets of patterns.

Through analysis of all NCEP variables mentioned in Section 3.1, we selected 2 variables to define the circulation patterns: 850 mb geopotential height ( $Z_{g850}$ ) and 850 mb specific humidity ( $Q_{850}$ ). These 2 variables produce a good overview of global atmospheric conditions (note that they are not identical to the variables used in  $\mathbf{X}_t$  to determine the transition probabilities between the different states). We focused on the region bounded by  $26\text{--}71^\circ\text{N} \times 115\text{--}61^\circ\text{W}$ , with 340 grid cells covering North America from the Rocky Mountains to the East Coast and from the Mexican border to the Arctic Circle. Of the 680 principal components (or EOFs), 100 (representing  $>99\%$  of the variance) were retained to run the EM algorithm. Amongst all tested variance structures and all numbers of patterns, the optimal BIC was obtained for  $K = 3$  clusters with diagonal variance matrix and varying volumes.

The clusters for geopotential heights (Fig. 4) represent alternating lower and higher pressures over Hudson Bay, with Pattern 1 corresponding to climatological mean winter heights of higher pressure systems over the Great Basin and lower pressures to the Northeast. The humidity patterns are more variable, with (Pattern 1) large positive anomalies centered inland from

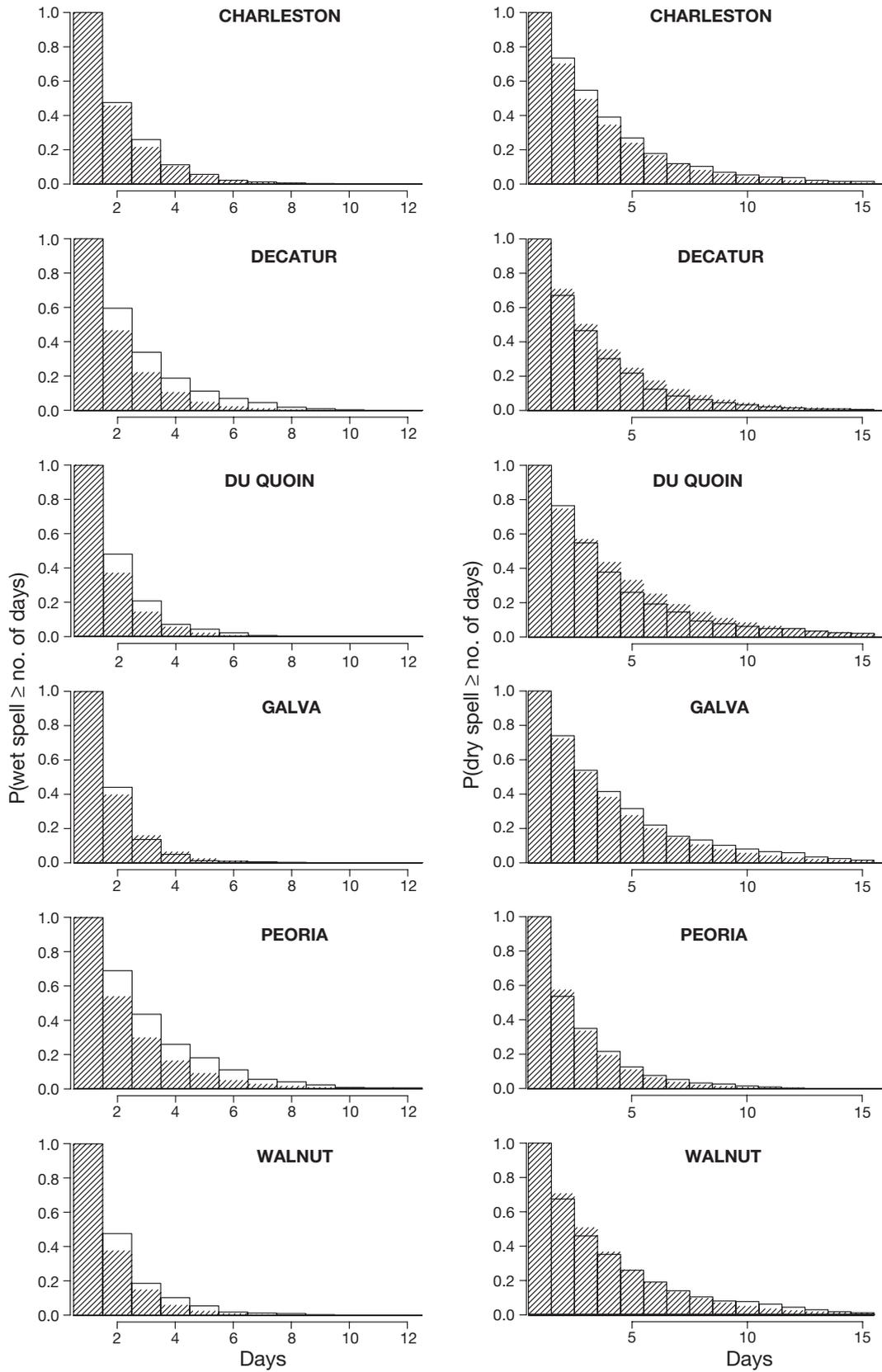


Fig. 3. Probabilities of duration of wet (left) and dry (right) spells for precipitation pattern-based simulated (hatched) and observed (clear) precipitation for 1952–1971

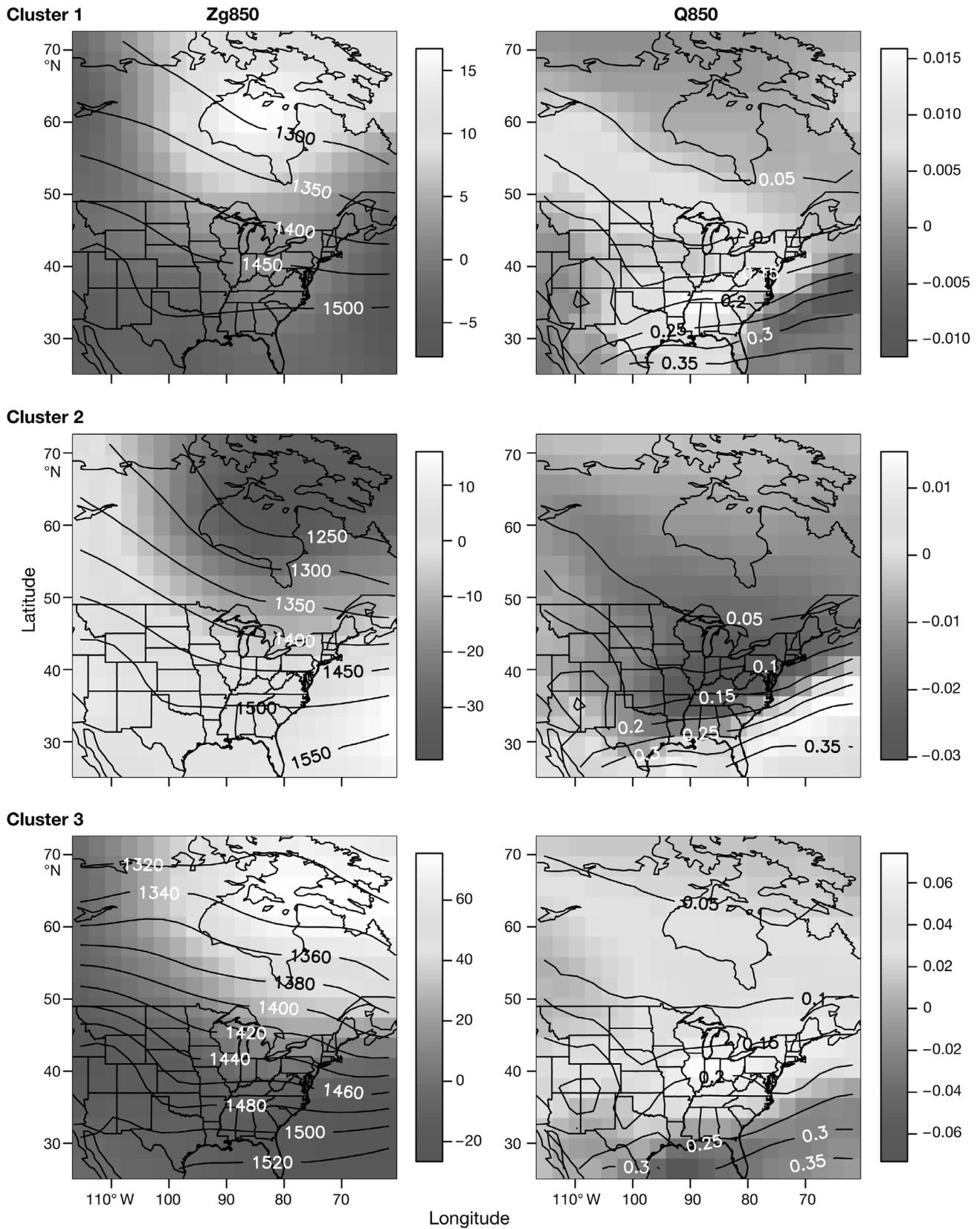


Fig. 4. Circulation patterns determined by the EM algorithm on geopotential height  $Zg_{850}$  and specific humidity  $Q_{850}$ . Anomalies are indicated by shading, and geopotential and specific humidity composites by isolines

the Gulf Coast, (Pattern 2) seaward from the Gulf and (Pattern 3) Atlantic coasts, and over Illinois and the eastern Midwest.

Fitting the NSWT model to the circulation patterns was performed in the same way as for the precipitation patterns (Section 2.2). Using the parameters obtained from the fitting period (1980–1999), simulations were performed for the evaluation period, as described in Section 3.2, but on the basis of upper-air circulation patterns, instead of surface precipitation patterns.

We again used QQplots to evaluate the quality of these simulations (Fig. 5). Similarly, circulation-based NSWT-simulated wet and dry spells are compared to observations in Fig. 6, which can again be directly contrasted with the comparison of precipitation-based NSWT simulations vs. observations in Fig. 3. For most of the stations, the QQplots obtained from the approach based on precipitation patterns (Fig. 2) are closer to observed data than those obtained from the approach fitted to circulation patterns (Fig. 5). In most cases, the circulation pattern-based approach exhibits a consistent over-estimation of precipitation quantiles. Although both approaches (i.e. NSWT simulations fitted to both precipitation and circulation patterns) result in relatively similar QQplots for some sites (e.g. Peoria), for most stations the precipitation-patterns

approach produces daily winter precipitation quantiles that are clearly closer to observations than those produced by the circulation-patterns approach.

Equally, the precipitation-based probabilities for wet and dry spells (Fig. 3) are better than those simulated from circulation patterns (Fig. 6), which again display a general tendency to under-estimate the occurrence of both wet and dry spells, relative to both observed and precipitation pattern-based simulations.

Given the systematic bias in the circulation pattern-driven approach, we hypothesize that perhaps the NSWT method itself may not be suited to circulation patterns or to the predictors used, or the statistical relationships between the larger region used to determine the circulation patterns and the state of Illinois might be difficult to capture. However, these hypotheses remain to be verified.

#### 4. CONCLUSIONS

This study describes the development of a relatively simple and transparent stochastic model for downscaling precipitation intensities, and evaluates this model based on observed precipitation for 37 stations located in the state of Illinois. The downscaling method, based

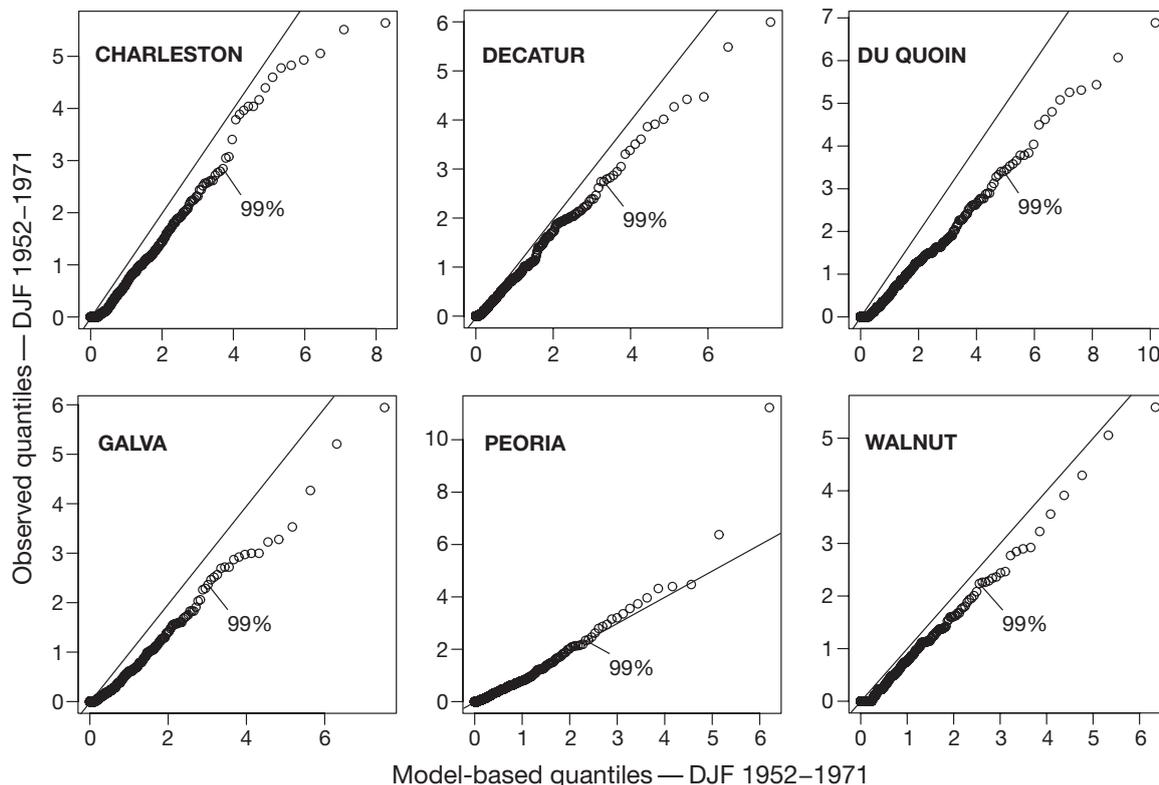


Fig. 5. QQplots of precipitation simulated from 3 circulation patterns vs. observed precipitation (in cm) for Dec, Jan, Feb 1952–1971; 99th percentiles are indicated

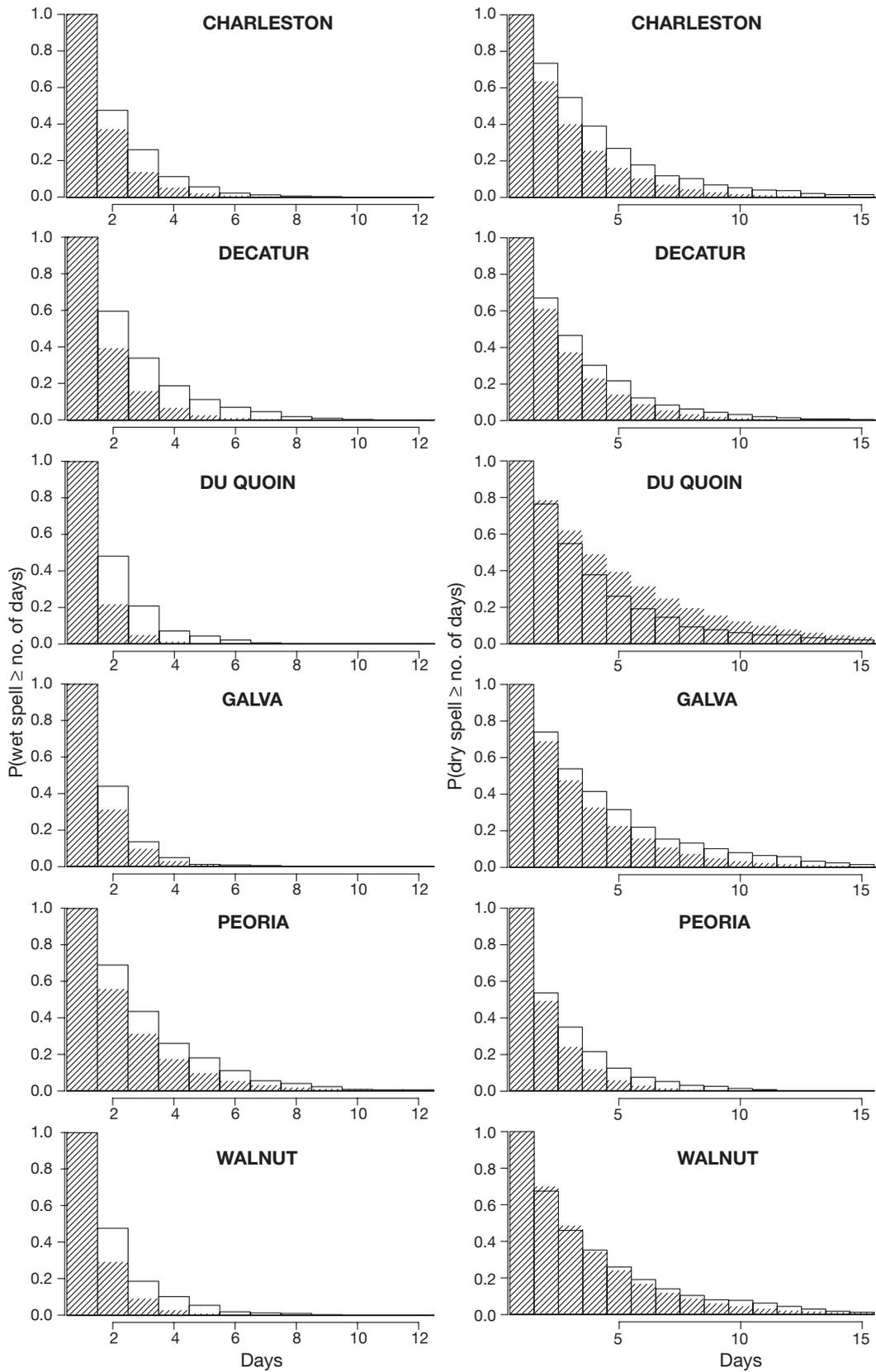


Fig. 6. Probabilities of duration of wet (left) and dry (right) spells for circulation pattern-based simulated (hatched) and observed (clear) precipitation for 1952–1971

on a weather typing approach, uses a nonhomogeneous Markov model (NMM) to represent the transition process from one weather pattern to another, here driven by upper-air atmospheric variables  $Zg_{850}$ ,  $Q_{850}$  and  $\Delta Td_{850}$ . In contrast to the nonhomogeneous hidden Markov model proposed by Bellone et al. (2000), our NMM is not hidden. Instead, we defined our patterns through a clustering step prior to the parameter estimations. Moreover, local rainfall occurrence probabilities were not held constant, but are stochastically influenced by large-scale atmospheric variables through logistic regression models with weather state- and site-specific parameters.

For comparison purposes, we defined 2 distinctly different types of weather patterns, based on (1) upper-air circulation fields, and (2) surface precipitation. The more traditional circulation patterns were defined here by a model-based clustering method through a mixture of Gaussian distributions. This method gives consistent atmospheric structures which, when used to fit the NSWTT downscaling approach, resulted in relatively successful simulation of local precipitation, although with a distinct tendency to over-estimate precipitation amounts, especially for high quantiles. However, when the NSWTT method was fitted to precipitation patterns that are directly related to the observed rainfall values, much more accurate simulations and characteristics of observed rainfall were derived, as evaluated through QQplots and probabilities of the length of wet/dry spells. In particular, quantile–quantile plots between observed and precipitation pattern-based simulated precipitation were nearly 1:1 up to the 99th percentile.

We therefore conclude that, for a given downscaling approach such as the NSWTT method, the large-scale patterns used to fit the model to observations are key in determining the accuracy of the method. Specifically, a precipitation pattern-based approach was better than the more traditional circulation pattern-based approach, particularly for higher precipitation quantiles. Furthermore, our goal of developing a model that was conceptually different from and potentially more transparent than an NHMM was also achieved through the nonhomogeneous stochastic weather typing (NSWT) method described here, which incorporated a non-hidden NMM in a more 'classical' weather typing context.

## 5. NEXT STEPS

Despite the initial success of this method, there are still a number of possible improvements. For example, the precipitation pattern-based QQplots in Fig. 2 show that, although the core of the precipitation is usually well-simulated, the highest values (>99%) can be either under- (e.g. Peoria) or over-estimated (e.g. Du

Quoin). Extreme precipitation events often have significant economic impacts, and improving simulation of these events is crucial.

We expect that introducing spatial dependence into the NSWTT scheme would also improve the correlation between precipitation events at individual stations. As a first step, we could model only the dependence of rainfall occurrence between the different sites, while intensity remained independent. As a second step, the dependence for both occurrence and intensity could be modeled simultaneously. Since this approach could considerably complicate the estimation step and slow down the process of estimation, an alternative could be to consider modern statistical methods of resampling such as bootstrapping (Efron & Tibshirani 1993). Resampling methods in which entire precipitation fields are resampled simultaneously present the advantage of retaining the spatial structure of the data without adding any new modeling.

It would also be valuable to know how our NSWTT approach performs compared to a NHMM approach and to a weather typing approach that does not incorporate the Markov assumption. Conditioning transition pattern processes and rainfall occurrence probabilities on large-scale atmospheric variables  $\mathbf{X}_t$  implies that these 2 features will be altered if  $\mathbf{X}_t$  fields change (see Section 2). For this reason, we expect that the accuracy of local precipitation projections under conditions of global climate change affecting the atmospheric variables  $\mathbf{X}_t$  would be greater than methods that did not include a dependence on large-scale circulation fields. However, many weather typing or, more generally, statistical downscaling methods without a Markov assumption also force the pattern frequencies or the rainfall probabilities to evolve with atmospheric variables.

In terms of future applications, the success of this method in simulating 1952–1971 observed precipitation suggests that this method may be reasonably applied to climate projections from an atmosphere–ocean general circulation model (AOGCM) to both supplement and potentially replace costly regional climate model (RCM) simulations in projecting future changes in precipitation. However, before applying this or any method to future projections, it is important to evaluate the ability of AOGCMs to reproduce the observation or reanalysis-based patterns used to fit the downscaling method. Furthermore, a fundamental assumption in any statistical downscaling approach is that the relationship between independent and dependent variables (here: weather states and local climate variables) will continue to be valid under future forcing. Some recent works suggest this assumption may be reasonable (e.g. Wood et al. 2004). Combining a statistical approach with regional model simulations

would enable us to largely address this issue and thus remove one of the primary caveats of the statistical vs. the dynamical approach.

Validation of a statistical downscaling approach such as that presented here could be done in 2 steps (see Charles et al. 1999): (1) Fitting our statistical model on AOGCM and RCM outputs for a present-day time period; the AOGCM data would provide us with simulated large-scale variables, while the RCM outputs would be used as local-scale 'observed' data. (2) AOGCM output characterizing a future time period, for which we have the corresponding AOGCM-forced RCM output, could be used to verify if the simulated future rainfall based on AOGCM output alone corresponds well with RCM-simulated precipitation; this step could be applied to various scenarios, to compare the validity of the simulation under climate change conditions.

Although some limitations can arise from RCM biases and caveats in their parameterization of small-scale physical processes, this validation would provide us with key information regarding the extent to which circulation and precipitation pattern-based statistical downscaling can be relied on to simulate the regional to local-scale impacts of future change. Once validated, this approach could be used to generate high-resolution and site-specific projections of future climate change at low computation cost, compared with RCM simulations. These projections could be used as input for impact assessments, where the effect of simulated changes in precipitation distribution, frequency and intensity on agricultural crops, water resources, flooding, etc. could be evaluated, also with regard to economic consequences (e.g. Hayhoe et al. 2004).

*Acknowledgements.* Although this research has been funded wholly or in part by the United States Environmental Protection Agency through STAR Cooperative Agreement #R-82940201 to the University of Chicago, it has not been subjected to USEPA peer and policy review; it therefore does not necessarily reflect the views of USEPA, and no official endorsement should be inferred.

#### LITERATURE CITED

- Akkur N, Cameron I, Auliciems A, Verrall K (1992) Synoptic ozone risks at Brisbane, Australia—Prediction using multivariate analysis and rule induction. In: Proc Int Clean Air Conf, Brisbane
- Bárdossy A, Muster H, Duckstein L, Bogardi I (1993) Knowledge based classification of circulation patterns for stochastic precipitation modelling. In: Proc Int Conf Stochastic Methods Hydrol Environ Eng, Waterloo, ON, p 201–215
- Bellone E, Hughes J, Guttorp JP (2000) A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Clim Res* 15:1–12
- Bischoff S, Vargas W (2003) The 500 and 1000 hPa weather type circulations and their relationship with some extreme climatic conditions over southern South America. *Int J Climatol* 23:541–556
- Charles SP, Bates B, Whetton P, Hughes J (1999) Validation of downscaling models for changed climate conditions: case study of southwestern Australia. *Clim Res* 12:1–14
- Davis R, Kalkstein L (1990) Using a spatial synoptic climatological classification to assess changes in atmospheric pollution concentration. *Phys Geog* 11:320–342
- Davis R, Dolan R, Demme G (1993) Synoptic climatology of Atlantic coast northeasters. *Int J Climatol* 13:171–189
- Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 39:1–38
- Diday E, Vrac M (2005) Mixture decomposition of distributions by copulas in the symbolic data analysis framework. *Discrete Appl Math* 147:27–41
- Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman & Hall, London
- Fraley C, Raftery A (2002) Model-based clustering, discriminant analysis and density estimation. *J Am Stat Assoc* 97:611–631
- Hayhoe K, Cayan D, Field C, Frumhoff P and 15 others (2004) Emission pathways, climate change, and impacts on California. *Proc Natl Acad Sci USA* 101:12422–12427
- Hewitson B, Crane R (2002) Self organizing maps: applications to synoptic climatology. *Clim Res* 26(10):1315–1337
- Hughes J, Guttorp P (1994) A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resour Res* 30:1535–1546
- Hughes J, Guttorp P, Charles SP (1999) A non-homogeneous hidden Markov model for precipitation occurrence. *Appl Stat* 48:15–30
- Huth R (2001) Disaggregating climatic trends by classification of circulation patterns. *Int J Climatol* 21:135–153
- Jeffries N, Pfeiffer R (2000) A mixture model for the probability distribution of rain rate. *Environmetrics* 12:1–10
- Kalkstein LS, Corrigan P (1986) A synoptic climatological approach for geographical analysis: assessment of sulfur dioxide concentrations. *Ann Assoc Am Geogr* 76:381–395
- Lamb H (1972) British Isles weather types and a register of daily sequence of circulation patterns. *Geophys Mem* 116, HMSO, London
- McLachlan G, Peel D (2000) Finite mixture model. Wiley, New York
- Mearns L (2003) Issues in the impacts of climate variability and change on agriculture. *Clim Change* 60:1–216
- Mearns L, Bogardi I, Giorgi F, Matyasovszky I, Palecki M (1999) Comparison of climate change scenarios generated from regional climate model experiments and statistical downscaling. *J Geophys Res* 104:6603–6621
- Murphy J (1999) An evaluation of statistical and dynamical techniques for downscaling local climate. *J Clim* 12: 2256–2284
- Naveau P, Vrac M, Genton M, Chédin A, Diday E (2004) Two statistical methods for improving the analysis of large climatic data sets: general skewed Kalman filters and distributions of distributions. In: Sanchez-Vila X, Carrera J, Gomez-Hernandez J (eds) Quantitative geology and geostatistics, Vol. 13. Geostatistics for environmental applications. Kluwer Academic, Barcelona
- Pongracz R, Bartholy J, Bogardi I (2001) Fuzzy rule-based prediction of monthly precipitation. *Phys Chem Earth* 9: 663–667
- Schnur R, Lettenmaier D (1998) A case study of statistical

- downscaling in Australia using weather classification by recursive partitioning. *J Hydrol* 212–213:362–379
- Smyth P, Ide K, Ghil M (1999) Multiple regimes in Northern Hemisphere height fields via mixture model clustering. *J Atmos Sci* 56:3704–3723
- von Storch H, Zwiers F (1999) *Statistical analysis in climate research*. Cambridge University Press, Cambridge
- Vrac M, Chédin A, Diday E (2005) Clustering a global field of atmospheric profiles by mixture decomposition of copulas. *J Atmos Ocean Technol* :1445–1459
- Vrac M, Hayhoe K, Stein M (2007) Identification and inter-model comparison of seasonal circulation patterns over North America. *Int J Climatol* 27(5):603–620
- Ward J (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58:236–244
- Wilby R, Wigley T, Conway D, Jones P, Hewitson B, Main J, Wilks D (1998) Statistical downscaling of general circulation model output: a comparison of methods. *Water Resour Res* 34:2995–3008
- Wilby R, Dawson C, Barrow E (2002) SDSM— a decision support tool for the assesment of regional climate change impacts. *Environ Model Software* 17:147–159
- Wilks D (1999) Multisite downscaling of daily precipitation with a stochastic weather generator. *Clim Res* 11:125–136
- Wilks D, Wilby R (1999) The weather generation game: a review of stochastic weather models. *Prog Phys Geogr* 23: 329–357
- Wood A, Leung L, Sridhar V, Lettenmaier D (2004) Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Clim Change* 62: 189–216
- Zorita E, Hughes J, Lettenmaier D, von Storch H (1993) Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. Report 109, Max-Planck-Institut für Meteorologie, Hamburg

*Editorial responsibility: Bryson Bates,  
Wembley, Western Australia, Australia*

*Submitted: June 21, 2006; Accepted: June 11, 2007  
Proofs received from author(s): August 27, 2007*