A stochastic weather model for generating daily precipitation series at ungauged locations in the Catskill Mountain region of New York state

Myeong-Ho Yeo1 | Allan Frei1,2 | Rakesh K. Gelda3 | Emmet M. Owens3

1Institute for Sustainable Cities, Hunter College, New York
2Department of Geography, Hunter College, City University of New York, New York
3Water Quality Modeling Section, New York City Department of Environmental Protection Police, Kingston, New York

Correspondence
Myeong-Ho Yeo, 695 Park Avenue, New York, NY 10065.
Email: myeongho.yeo@gmail.com

Funding information
New York City Department of Environmental Protection

Abstract
Information on the variability of precipitation in time and space is critical for many water resource projects. However, precipitation records at the location of interest are often either limited or unavailable due to an inadequate network of rainfall measurements. To address this need, regionalization methods have been employed to characterize spatial patterns of precipitation and to transfer precipitation information from one location to another where records are scarce. Hence, the overall objective of the present paper is to propose a stochastic weather model for generating daily precipitation at ungauged locations. The proposed approach consists of two components: (a) a regionalization approach for identifying homogeneous groups of observed daily precipitation series, and (b) a stochastic model for constructing daily precipitation events at ungauged locations within homogeneous groups. This statistical approach identifies groups of precipitation stations with similar statistical characteristics based on the combination of two multivariate statistical techniques: principal component analysis (PCA) and ordinal factor analysis (OFA). While the application of PCA in climatological regionalization studies based on precipitation amount is common, the application of OFA to include precipitation occurrence in the identification of regions is unusual. The feasibility of the approach is assessed using daily precipitation data from a network of precipitation stations in the Catskill Mountain region of New York State, United States.

KEYWORDS
climate regionalization, estimation of daily precipitation, ordinal factor analysis, principal component analysis, ungauged precipitation

1 | INTRODUCTION

The management and allocation of water resources have been considered as one of the most significant endeavours in human society because water plays a vital role in all natural and environmental systems. In addition, according to the Intergovernmental Panel on Climate Change (IPCC) Synthesis Report (IPCC, 2014), climate change has had significant impacts on precipitation processes in many regions of the world, including increasing trends in precipitation amounts in North America, northern Europe, and northern and central Asia. The northeastern United States, where our study region...
is located, has experienced a larger increase in extreme precipitation than any other region in the United States (Rosenzweig, 2011; Melillo et al., 2014). It is thus essential to consider the potential impacts of climate change on water resources. Information on the variation in time and space of precipitation and the resulting runoff is essential for these climate change impact assessment studies. However, many watersheds do not have an adequate network of precipitation measurements to provide a sufficient record of observed data.

Because the Catskill Mountains provide most of the water supply for New York City (NYC) and other regional municipalities, changes to the hydrological cycle can affect, in addition to several hundred thousand local residents, over 9 million residents of New York State (DEP, 2017). The geographical and meteorological features of the region result in complex seasonal and spatial patterns of precipitation (see section 3). In the context of regional impact studies, information on the spatial distribution of precipitation at a sufficiently high resolution is essential for accurate calculation of local values of hydrologic variables such as streamflow (Baeriswyl and Rebetez, 1997). Due to the high spatial variability of precipitation and low density of measurement stations in mountainous areas, the estimation and prediction of hydrologic variables with climate change conditions remains a challenge (Sivapalan, 2003; Yeo, 2014).

To address this need, regionalization methods have been employed to understand the spatial behaviour of precipitation in order to effectively transfer precipitation information from a location with sufficient observations to another where available records are scarce (Baeriswyl and Rebetez, 1997; Nguyen et al., 2002). More specifically, these methods have been developed and employed according to two main objectives: characterizing spatial dependency (homogeneity) and reducing uncertainty in the modeling of precipitation at different locations. The implication of spatial dependency is that the precipitation records from two or more stations in a delineated homogenous group have similar statistical characteristics. González and Valdés (2008) stress that regionalization is a powerful tool for improving the accuracy of precipitation estimation at a location of interest. Consequently, evaluation of the similarity of observed precipitation series at different locations is a first step in reducing uncertainty, leading to the reliable and accurate estimation of precipitation over an area or over multiple sites (González and Valdés, 2008).

Because it may be problematic in hydroclimatological studies to delineate regions over data-sparse areas using only in-situ precipitation records, some studies consider the inclusion of large-scale circulation patterns to help construct meaningful homogeneous regions (Satyanarayana and Srinivas, 2008; Satyanarayana and Srinivas, 2011; Asong et al., 2015). To estimate the effects of large-scale patterns to local weather conditions, it is essential to use long-term weather information rather than short-term data in order to avoid noise resulting from local weather variability. The previous studies use monthly time series of historical observations and atmospheric re-analysis from National Centers for Environmental Prediction (NCEP) for describing the effect of large scale atmospheric circulation on local weather condition. However, at relatively small spatial scales, such as our study area, local geographic features such as elevation, wind direction, and aspect play the most significant roles of forming spatial patterns of weather conditions (Burns, 1953; Daly et al., 1994; Daly et al., 2008). Therefore, in this study we evaluate a regionalization method that includes statistics derived from in-situ daily precipitation measurements, but no large scale circulation features.

Many regionalization methods determine the statistical similarity of precipitation at different locations based on the correlation of precipitation amount. A limitation of these conventional approaches is that the resulting regions do not include information on spatial variation in precipitation occurrence (wet/dry-day). In this study we propose a regionalization technique based on the similarities of both precipitation amount and occurrence at different locations using the combination of principal component analysis (PCA) and ordinal factor analysis (OFA). These statistical techniques identify groups of stations with similar precipitation statistics based exclusively on the station records: the spatial location of stations is not included in the analysis. The reason that they are sometimes referred to as “regionalization” techniques is that, due to the physics of precipitation processes, stations with similar statistical properties often naturally fall into spatially contiguous regions.

In addition to the regionalization, stochastic weather generators (SWGs) have been used to generate synthesized hydrometeorologic information in agricultural, hydrological, or climate change impacts studies (Wilks and Wilby, 1999; Semenov and Barrow 2002). Schuurmans and Bierkens (2006) address the spatial variability of daily precipitation as a major contribution to discharge, groundwater level and soil moisture. A number of studies on multi-site (or distributed) SWGs have also been proposed (Wilks, 1998; Wilby et al., 2003; Khalili et al., 2009; Asong et al., 2016; Peleg et al., 2017; Evin et al., 2018). Three main components of multi-site SWGs are the representation of precipitation amount, of precipitation occurrence, and of spatial variability. In general, major uncertainty in models of precipitation amount is to select an appropriate distribution to account for statistics of the observed variables. This is because the data pool is not a population but a sample. Once a suitable distribution is chosen, the choice of parameterization method adds additional uncertainty. Moreover, Markov Chain Models have been used with different order states to derive the random feature of wet/dry process. Although these SWGs provide synthesized information on daily time scales, and model
statistics match the observed statistics (Wilks and Wilby, 1999), there is no match between the daily observed and estimated data. Our study therefore proposes a stochastic model to transfer daily precipitation information from a station to another based on actual observed daily data rather than the long-term observed data pool. Moreover, applying precipitation signals from outside the homogenous regime to the SWGs is likely to add more noise or uncertainty (Schuurmans and Bierkens, 2006). It could imply that the delineation of statistically homogenous weather region is an essential step prior to weather generation in order to represent spatial variability and accurate characteristics. Hence, the weather modelling in this study is conducted using the rain gauges within the identified homogeneous regions.

The remainder of this paper is organized as follows: section 2 describes the regionalization approach using OFA, section 3 describes the proposed stochastic model for estimating daily precipitation series at ungauged locations with our study region and data. In section 4, we present our calibration and validation analyses. Finally, our conclusions are presented in section 5.

2 | REGIONALIZATION USING OFA

Regionalization is more complicated for daily precipitation than for most other meteorological variables because it is a two-dimensional vector composed of both a discrete (i.e., occurrence) and highly skewed continuous (i.e., amount) components. In terms of regionalization of daily precipitation, while there is no agreement about which property is more important, most existing regionalization methods (see Figures S1 and S2 and Tables S1 and S2) consider only the correlation of precipitation amounts. Precipitation occurrence may however be important in regions such as the Catskill Mountains region where orography modulates precipitation patterns.

To explicitly include precipitation occurrence in a regionalization analysis, Yeo (2014) took the novel approach of applying a combination of PCA and OFA to precipitation observations from the southern part of Korean Peninsula, where two complex mountain systems play an important role modulating precipitation occurrence and amount. His study indicated that the inclusion of precipitation occurrence using OFA adds spatial information that is unavailable from PCA alone and improves simulation results generating ungauged daily precipitation series. As an extension of the unpublished study of Yeo (2014), here we apply a modified version of the Yeo (2014) approach to historical precipitation observations from the Catskill Mountains region, and evaluate whether the inclusion of occurrence information through OFA improves simulation results. Table 1 summarizes the statistical techniques that have been used to identify factors for categorical and interval data (Bartholomew et al., 2011).

3 | STUDY REGION, DATA, AND METHODOLOGY

3.1 | Study area

The Catskill Mountains, located in southeastern New York State, between latitude 41.78° and 42.35°N and longitude 75.12° and 74.13°W (Figure 1), are the home of six reservoirs and associated watersheds that historically provide approximately 90% of the drinking water to NYC. This regionalization study encompasses these watersheds and the surrounding region of about 31,000 km² in New York and Pennsylvania. As a part of the Appalachian Mountains, numerous peaks in this region are over 900 m above sea level. The complex topography includes ridges oriented from the southwest to northeast, as well as a southeast-to-northwest oriented escarpment defining the northeastern boundary of the region. Mean seasonal precipitation (Figure 2) is greater along the southeastern escarpment where coastal storms laden with tropical moisture provide much of the precipitation, while mid-latitude storms with various tracks are also important in this region (Towey et al., 2018).

3.2 | Data used for calibration and validation of the regionalization technique

Historical precipitation data from 106 rain gauge stations (Table 2) within this region were obtained from the Northeast Regional Climate Center (NRCC) and National Climatic Data Center (NCDC) of National Oceanic and Atmospheric Administration (NOAA). As shown in Figure 3, the number of stations measuring daily precipitations over the study area has

---

**Table 1** Classification of latent variable methods according to data types

| Latent variables | Continuous | Categorical |
|------------------|------------|------------|
| Manifest variables | Factor analysis | Latent trait analysis* |
| Latent profile analysis | Latent class analysis |

*Ordinal factor analysis technique is one type of latent trait analysis.
**FIGURE 1** The study area encompasses the reservoir watersheds, located West-Of-Hudson, to supply drinking water to the New York City. Points denote the rain-gauge stations used in this study.

**FIGURE 2** Interpolated seasonal mean precipitation (mm) for spring season in study area.
| Site code | Station                   | Start year | End year | Latitude | Longitude |
|----------|---------------------------|------------|----------|----------|-----------|
| S1       | ALBANY                    | 1941       | 1981     | 42.65    | -73.75    |
| S2       | ALBANY INTERNATIONAL AIRPORT | 1950       | 2010     | 42.75    | -73.80    |
| S3       | ALCOVE DAM                | 1950       | 2010     | 42.47    | -73.93    |
| S4       | ANDES                     | 1950       | 1959     | 42.20    | -74.78    |
| S5       | ARENA                     | 1950       | 1970     | 42.10    | -74.72    |
| S6       | ARKVILLE                  | 1950       | 2010     | 42.14    | -74.65    |
| S7       | BAINBRIDGE 2 E            | 1941       | 1992     | 42.28    | -75.45    |
| S8       | BALSAM LAKE               | 1950       | 1969     | 42.03    | -74.60    |
| S9       | BERNE 2 S                 | 1967       | 1982     | 42.60    | -74.15    |
| S10      | BINGHAMTON                | 1951       | 2016     | 42.11    | -75.89    |
| S11      | BOVINA                    | 1950       | 1959     | 42.27    | -74.73    |
| S12      | BULLS BRG DAM             | 1948       | 2000     | 41.68    | -73.51    |
| S13      | BUSHNELLSVILLE 1 NE      | 1950       | 1959     | 42.17    | -74.40    |
| S14      | BUTTERNUT BROOK           | 1950       | 1959     | 41.92    | -74.67    |
| S15      | CAIRO 4 NW                | 1979       | 2010     | 42.32    | -74.03    |
| S16      | CALLICOON                 | 1950       | 1966     | 41.77    | -75.05    |
| S17      | CATSKILL 1 N              | 1964       | 1977     | 42.23    | -73.87    |
| S18      | CHERRY VALLEY 2 NNE       | 1950       | 2009     | 42.82    | -74.74    |
| S19      | CHINA 2                   | 1950       | 1970     | 42.10    | -75.38    |
| S20      | CLARYVILLE                | 1950       | 1967     | 41.91    | -74.57    |
| S21      | COBLESKILL                | 1956       | 1985     | 42.68    | -74.48    |
| S22      | COLD BROOK                | 1950       | 1959     | 42.02    | -74.27    |
| S23      | COOPERSTOWN               | 1950       | 2010     | 42.72    | -74.93    |
| S24      | CRAIGIE CLAIR             | 1950       | 1959     | 41.97    | -74.87    |
| S25      | DELHI 2 SE                | 1950       | 2010     | 42.25    | -74.91    |
| S26      | DEPOSIT                   | 1962       | 2012     | 42.07    | -75.43    |
| S27      | DIXON                     | 1949       | 1969     | 41.57    | -75.90    |
| S28      | DOWNSVILLE DAM            | 1959       | 2010     | 42.08    | -74.98    |
| S29      | EAST JEWETT               | 1950       | 2010     | 42.25    | -74.18    |
| S30      | EDGEWOOD                  | 1950       | 1963     | 42.13    | -74.23    |
| S31      | ELKA PARK                 | 1950       | 1959     | 42.17    | -74.17    |
| S32      | ELLENVILLE                | 1945       | 2014     | 41.72    | -74.40    |
| S33      | EQUINUNK                  | 1945       | 2001     | 41.87    | -75.27    |
| S34      | FISHS EDDY                | 1953       | 2007     | 41.97    | -75.18    |
| S35      | FREEHOLD 2 NW             | 1964       | 1978     | 42.37    | -74.07    |
| S36      | FROST VALLEY              | 1950       | 1980     | 41.97    | -74.55    |
| S37      | GARDINER 1 W              | 1957       | 1987     | 41.68    | -74.15    |
| S38      | GLENFORD                  | 1950       | 1959     | 42.02    | -74.13    |
| S39      | GRAHAMSVILLE              | 1950       | 2002     | 41.85    | -74.53    |
| S40      | GRAND GORGE               | 1950       | 1977     | 42.37    | -74.48    |
| S41      | GREENE                    | 1941       | 2001     | 42.32    | -75.77    |
| S42      | HALCOTT 2 NNE             | 1948       | 1959     | 42.22    | -74.47    |

(Continues)
| Site code | Station       | Start year | End year | Latitude | Longitude |
|-----------|---------------|------------|----------|----------|-----------|
| S43       | HAMILTON      | 1941       | 1963     | 42.82    | −75.53    |
| S44       | HARVARD       | 1950       | 1959     | 42.02    | −75.12    |
| S45       | HAWLEY 4 SW   | 1941       | 1972     | 41.45    | −75.27    |
| S46       | HIGH FALLS    | 1950       | 1967     | 41.83    | −74.13    |
| S47       | HIGHMOUNT     | 1950       | 1976     | 42.13    | −74.48    |
| S48       | KORTRIGHT 2   | 1960       | 1994     | 42.42    | −74.80    |
| S49       | LACKAWACK     | 1950       | 1960     | 41.78    | −74.40    |
| S50       | LAKE DELAWARE | 1950       | 1980     | 42.25    | −74.90    |
| S51       | LAKE HILL     | 1950       | 1959     | 42.07    | −74.18    |
| S52       | LANSING MANOR | 1986       | 2010     | 42.45    | −74.47    |
| S53       | LEWBEACH      | 1950       | 1959     | 41.95    | −74.83    |
| S54       | LEXINGTON 1 SE| 1950       | 1959     | 42.23    | −74.33    |
| S55       | LIBERTY 1 NE  | 1950       | 2010     | 41.80    | −74.74    |
| S56       | MANORKILL     | 1950       | 1994     | 42.38    | −74.32    |
| S57       | MARY SMITH    | 1950       | 1976     | 42.05    | −74.82    |
| S58       | MARYLAND 6 SW | 1984       | 2010     | 42.47    | −75.01    |
| S59       | MERRIMAN DAM  | 1962       | 2004     | 41.80    | −74.43    |
| S60       | MIDDLEBURG    | 1950       | 1976     | 42.60    | −74.32    |
| S61       | MILANVILLE    | 1945       | 1975     | 41.67    | −75.06    |
| S62       | MOHONK LAKE   | 1950       | 2010     | 41.77    | −74.16    |
| S63       | MONGAUP VALLEY 4 SSW | 1950 | 1973 | 41.57 | −74.79 |
| S64       | NARROWSBURG 4 SE | 1957 | 1998 | 41.57 | −75.02 |
| S65       | NEVERSINK     | 1950       | 1959     | 41.83    | −74.65    |
| S66       | NEW BERLIN    | 1941       | 1995     | 42.62    | −75.33    |
| S67       | NEW KINGSTON  | 1950       | 1984     | 42.23    | −74.70    |
| S68       | NEW PALTZ 4 SW| 1972       | 1982     | 41.68    | −74.13    |
| S69       | OAK HILL      | 1950       | 1961     | 42.40    | −74.15    |
| S70       | ONEONTA 3 SE  | 1950       | 1969     | 42.45    | −75.00    |
| S71       | ONEONTA STATE UNIVERSITY | 1971 | 1983 | 42.47 | −75.07 |
| S72       | PARKSTON      | 1950       | 1969     | 41.90    | −74.82    |
| S73       | PEEKAMOOST    | 1950       | 1963     | 41.93    | −74.38    |
| S74       | PHOENICIA     | 1950       | 2000     | 42.09    | −74.31    |
| S75       | PLEASANT MT 1 W | 1941 | 2001 | 41.74 | −75.45 |
| S76       | PORT JERVIS   | 1941       | 2001     | 41.39    | −74.69    |
| S77       | POUGHKEEPESIE | 1941       | 1970     | 41.68    | −73.93    |
| S78       | PRATTSVILLE   | 1950       | 2001     | 42.33    | −74.44    |
| S79       | PRESTON HOLLOW| 1950       | 1962     | 42.43    | −74.20    |
| S80       | RELAY         | 1950       | 1959     | 42.32    | −74.63    |
| S81       | RIPTON 1 N    | 1950       | 1967     | 41.85    | −74.05    |
| S82       | ROCK HILL 3 SW| 1957       | 2010     | 41.60    | −74.60    |
| S83       | ROCKDALE      | 1950       | 2007     | 42.38    | −75.40    |
| S84       | ROSCOE        | 1972       | 1979     | 41.93    | −74.90    |

(Continues)
decreased significantly since 1960. This decrease is particularly evident within the water supply region (bold boundary shown in Figure 1) where the number of stations dropped from 80 stations in the 1950s to 38 stations in the 1980s. Based on these changes in data availability, the period 1949–1959 was chosen to calibrate this regionalization analysis in order to ensure that the regionalization study includes as much spatial detail and as many stations as possible. Daily precipitation time series from 80 stations that do not have any periods of missing data longer than 2 months are used for calibration.

| Site code | Station                  | Start year | End year | Latitude | Longitude |
|-----------|--------------------------|------------|----------|----------|-----------|
| S85       | ROSENDALE 2 E            | 1957       | 2010     | 41.85    | −74.05    |
| S86       | ROXBURY                  | 1950       | 1971     | 42.28    | −74.57    |
| S87       | SCHENECTADY               | 1943       | 1984     | 42.80    | −73.92    |
| S88       | SEAGER                   | 1950       | 1970     | 42.05    | −74.55    |
| S89       | SHOKAN BROWN STATION      | 1953       | 2007     | 41.95    | −74.20    |
| S90       | SLIDE MOUNTAIN           | 1950       | 2010     | 42.02    | −74.42    |
| S91       | STAMFORD                 | 1950       | 2003     | 42.40    | −74.63    |
| S92       | STEWART FLD              | 1948       | 1969     | 41.50    | −74.10    |
| S93       | SUNDOWN                  | 1950       | 1959     | 41.88    | −74.47    |
| S94       | SUSQUEHANNA              | 1941       | 2001     | 41.95    | −75.60    |
| S95       | TANNERSVILLE 2 E         | 1960       | 1974     | 42.20    | −74.10    |
| S96       | TERRY CLOVE              | 1950       | 1959     | 42.13    | −74.90    |
| S97       | UNADILLA                 | 1950       | 1978     | 42.32    | −75.32    |
| S98       | UNADILLA 2 N             | 1976       | 2010     | 42.35    | −75.32    |
| S99       | WALDEN 2 NE              | 1941       | 1959     | 41.57    | −74.17    |
| S100      | WALTON                   | 1957       | 1997     | 42.17    | −75.13    |
| S101      | WEST KILL                | 1950       | 1959     | 42.20    | −74.38    |
| S102      | WEST_POINT               | 1941       | 2001     | 41.40    | −73.95    |
| S103      | WEST SHOKAN 3 SW         | 1950       | 1959     | 41.95    | −74.32    |
| S104      | WESTERLO                 | 1950       | 1959     | 42.52    | −74.05    |
| S105      | WHITNEY POINT LAKE       | 1941       | 2001     | 42.33    | −75.97    |
| S106      | WINDHAM                  | 1950       | 2008     | 42.30    | −74.20    |

**FIGURE 3** Data availability in this study. The period 1, in which the number of stations with measurements each year is over 95%, is from 1949 through 1959, and the period 2 is from 1981 through 1991.
To validate the regionalization results, the same analysis is performed for the interval 1981–1991, when fewer stations (38 stations) are available. A comparison of the regionalization analyses conducted during the calibration (1950s) and validation (1980s) periods is used to indicate whether the delineated climatic regions are dependent on time (The validation methodologies for the stochastic weather model in combination with the regionalization results are described in sections 3.5 and 3.6).

### 3.3 Regionalization

In the implementation described here, PCA and OFA are employed sequentially in order to identify homogeneous groups of daily precipitation stations. PCA is performed first to identify "groups" with correlated precipitation amount. The number of principal components (PCs) retained for further analysis is based on Kaiser's rule (Kaiser, 1958), which includes only PCs with eigenvalues greater than 1. OFA is then performed on each PCA group independently, resulting in a set of OFA factors and loadings. The factors are then rotated using PROMAX (Hendrickson and White, 1964), an oblique rotation that maintains the maximum differences, but not necessarily the orthogonality, between the rotated factors, and provides more robust results than orthogonal rotation methods (White et al., 1991). As a result of the combined PCA and OFA analyses, each station is placed in one group with other stations that have similar statistical properties for precipitation occurrence and amount.

### 3.4 A stochastic precipitation model for Ungauged locations

Data from stations within each region are used to estimate precipitation at other stations, or at ungauged locations, within the region, including probabilistic information derived from using a stochastic model. The stochastic modelling of the precipitation process is based on the combination of two different components: the modelling of precipitation occurrence and of precipitation amount.

#### 3.4.1 Precipitation occurrence model

The modelling of precipitation occurrence is based on the homogenous grouping given by the application of the PCA/OFA regionalization method. Let \( F_j \) be the factor score for a given day \( j \) as defined by:

\[
F_j = \left( \sum_{i=1}^{s} e_i O_{i,j} \right) / S
\]

where \( e_i \) is the factor loading associated with the rain gauged station \( i \) within an identified homogeneous region of \( s \) stations, \( O_{i,j} \) is the precipitation occurrence at a station \( i \) for day \( j \), and \( S \) is the number of stations used in the calculation procedure. The minimum threshold of 1 mm/day is used to define a series of \( O_{i,j} \). With the homogenous region identified by the OFA regionalization method, the factor score \( F_j \) represents the ratio of the number of stations in a given region where precipitation occurs on day \( j \) to all stations within the identified region: in other words, the probability of precipitation occurrence. A random number chosen from the uniform distribution \( r^o_j \left( 0 \leq r^o_j \leq 1 \right) \) is then used to determine precipitation occurrences for each day \( j \) by comparing the random number to the value of \( F_j \) as follows:

\[
\begin{align*}
\text{if } r^o_j \geq F_j, \text{ then wet at day } j \\
\text{if } r^o_j < F_j, \text{ then dry at day } j
\end{align*}
\]

#### 3.4.2 Precipitation amount model

Regarding the modelling of precipitation amount, a log-transformation technique is applied to the highly skewed precipitation distribution to more closely approximate a normal distribution (Hay et al., 1991):

\[
Y_{i,j} = \ln P_{i,j}
\]

where \( Y_{i,j} \) represents the corresponding log-transformed precipitation amount, and \( P_{i,j} \) denotes the precipitation amount at a station \( i \) and day \( j \). When a day is determined to be dry by comparing to the threshold (1 mm/day), the log-transformed amount at day \( j \) is set to zero in order to avoid negative and undefined (for \( P_{i,j} = 0 \)) values in the new transformed regime. In accordance with the log-transformation, the mean \((\bar{P})\) of \( P_{i,j} \) can be written by the mean \((\mu_Y)\) and the variance \((\sigma_Y^2)\) of \( Y_{i,j} \) as follows:

\[
E(P) = \bar{P} = \exp \left( \mu_Y + \frac{\sigma_Y^2}{2} \right)
\]

Thus, the regional expected value of the daily precipitation amount can be estimated by:

\[
E \left( P^R_j \right) = \bar{P}^R_j = W_j \times \exp \left( \mu^R_j \right)
\]

in which \( \bar{P}^R_j \) is the regional mean of the daily precipitations for day \( j \), \( \mu^R_j \) is the regional mean of the log-transformed daily precipitations for day \( j \), and \( W_j \) is the correction factor for matching the expected values of the original and log-transformed daily precipitations.
After the daily regional mean precipitation ($P_{ij}$) has been estimated, a daily precipitation amount is stochastically generated by multiplying a random scaling factor ($\theta$) and the regional mean. The random scaling factor was implemented for generating variance of daily precipitation amount using the random number sampled from a uniform distribution ($0 < r^* \leq 1$) (Hay et al., 1991; Wilby et al., 1994). Equations (1) and (5) can be used to generate missing precipitation series at a given site within a coherent region.

\[ P_j = P^*_j \times \theta = W_j \times \exp\left(\mu_{ij}^R\right) \times (-\ln (r^*)) \quad (5) \]

3.5 Cross-validation of the ungauged precipitation model during the calibration period

To cross-validate the stochastic model, daily precipitation data from stations in the identified homogeneous region during the calibration period (1950s) are used. Each station is deleted one at a time, and the other stations are used to estimate precipitation at the excluded station. After excluding each station, the PCA/OFA model is calibrated, and the daily precipitation series are estimated based on the data available at the remaining stations in the OR2. The computation procedure for estimating the daily precipitation series at an ungauged site can be summarized as follows:

1. Apply the PCA regionalization approach to the daily precipitation amount series.
2. Apply the OFA to each group identified by PCA to delineate the homogeneous regions.
3. Assume one rain gauge station in a homogenous region is an ungauged site. After excluding the selected rain gauge station data, calculate the tetrachoric correlation coefficients matrix using the daily precipitation occurrence data from the remaining rain gauge stations in the region.
4. Generate an ensemble of 20 stochastic realizations of daily precipitation occurrence series for an ungauged station based on the computed factor scores ($F_j$) (Equation (1)) and generated random number chosen from the uniform distribution ($r_j$).
5. Generate an ensemble of 20 realizations of daily precipitation amount series using the regional average of precipitation amount and the calculated weights (Equation (5)).
6. To keep precipitation rainfall only for the wet days, multiply the occurrence series and the amount series.

To evaluate the contribution of OFA to this weather generation procedure, three variations of this procedure were tested and compared to each other. In the first variation, referred to as Model-1, regions are defined by the application of PCA (i.e., OFA is completely excluded from the analysis). In other words, instead of applying OFA to the stations in Group 1, PCA is applied a second time to the stations in group 1, and then a stochastic weather model excluding the modelling part of daily occurrence is used. Model-2 is implemented by removing the second step completely, and calibrating the SWG on all stations in Group 1 to investigate the effect of heterogeneity on the weather generator. Model-3 is the full model described in the previous section, including both PCA and OFA.

3.6 Validation of the ungauged precipitation model

To validate the proposed weather model, we focus on one station, Grahamsville, because historical daily precipitation series at this station exists during both the calibration and validation periods. The computational procedure was carried out for both the calibration and validation periods. The computed precipitation series at Grahamsville are then statistically analysed and compared to the observed precipitation data using and number of metrics. First, precipitation occurrence is evaluated based on several indices: annual and monthly number of wet days; the Fraction Correct (FC) index, the Proportion of Wet-day Correct (PWC) known as the hit rate, and the Critical Success Index (CSI) (Wilks, 2011). The last three metrics are calculated based on the contingency table (Table 3) as follow:

\[ FC = \frac{a + d}{n} \]
\[ PWC = \frac{a}{a + c} \]
\[ CSI = \frac{a}{a + b + c} \quad (6) \]

FC is the fraction of the forecast correctly estimated, PWC is the expression of the probability of detection by the model, and CSI represents the proportion of correct forecasts to the total number of cases of precipitation either observed

| Station 1 | Wet | Dry |
|-----------|-----|-----|
| Wet       | a   | b   | a + b |
| Dry       | c   | d   | c + d |
|           | a + c | b + d | N = (a + b + c + d) |

TABLE 3 A contingency table to define a tetrachoric correlation coefficient between two stations
or forecast. Precipitation amount is evaluated based on annual and monthly mean and root-mean-square-error (RMSE).

4 | RESULTS

The regionalization analysis confirmed that, although the statistical techniques do not include any information about geographic location, the factor loadings are concentrated in particular regions. Such spatial clustering is reasonable based on the physics of precipitation: the prevailing winds and storm tracks combined with orography result in particular regional patterns of precipitation.

4.1 | Calibration of the regionalization model

Five PCs with eigenvalues greater than 1 explain 77.55% of the temporal variance in precipitation during the calibration period (1949–1959), and are therefore retained for further analysis (Table 4). In other words, the original time series from 80 stations has been reduced to only five components. PC5, which does not load onto a spatially coherent region and accounts for only ~1% of the total variance, is excluded from the remainder of the analysis. Figure 4 shows the rotated loading patterns corresponding to the remaining four extracted PCs. These loading patterns indicate that, while each component has some effect throughout the study area, each component is associated with a particular region. The first component has the largest factor loadings over the southeastern and eastern portions of the study area; component 2 loads most heavily over the northwestern corner of our region; component 3 loads most heavily over the northeastern escarpment; and component 4 over the southeastern and northeastern corners of the study region.

Based on these component loading patterns, we identify groups of stations that load most heavily on each component and fall within well-defined regions. Figure 5 shows the four groupings of stations identified by PCA. Group I, which includes stations that have their highest component loadings on PC1, represents a large part of Catskill Mountains including four of the six watersheds of the NYC water supply system (e.g., Ashokan, Neversink, Pepacton, and Rondout watersheds). The Cannonsville watershed and western Catskill Mountains region fall in Group II, while the northwestern Schoharie watershed is classified into Group III. Group IV includes stations located to the north and south of the NYC watersheds.

OFA was then applied within each of the four PCA regions. The results of the OFA analyses demonstrate that the spatial groupings identified by PCA based on precipitation amount can be further subdivided by OFA based on precipitation occurrence. OFA results are summarized in Figure 6. Eleven climatic regions (OR1 to OR11), where OR denotes a region defined by OFA, are identified. PCA Group I was further divided into three regions: OR1, covering the Pepacton watershed area and its higher elevation eastern ridges; OR2, covering the eastern/lower elevation Ashokan and Rondout basins; and OR3, covering the Neversink basin. OFA identifies two climatic regions within PCA Group II: within (OR4) and outside (OR5) of the Cannonsville watershed. Three distinct regions are found within PCA Group III: OR6 dominates most of the Schoharie watershed; OR7 dominates the signal mostly outside of the NYC watersheds; and OR8 dominates the lower elevation area (Grand Gorge and Manor Kill stations) of Schoharie and higher portion (Relay and Stamford stations) from Cannonsville. Finally, PCA Group IV is divided into three different sub-regions (OR9, OR10, and OR11) that lie outside of the watersheds. Results of the application indicate that the geographical features including elevation and the ridgelines (or divides) between watersheds play a role as demarcating regions and have a considerable influence on climatic patterns.

4.2 | Validation of the regionalization model

Figure 7 shows validation results based on application of the same procedure to daily precipitation data from the smaller group of 38 stations for 1981–1991 (methodology described in section 3.2). The similarity of regionalization results, despite the fact that they are based on fewer stations and performed with data from a different decade, suggests that the spatial pattern of regions of similar precipitation statistics is fairly constant through time, and is therefore largely controlled by topography (Baeriswyl and Rebetez, 1997; Comrie and Glenn, 1998).

4.3 | Cross-validation of the Ungauged precipitation model during the calibration period

As further validation of the model, the stochastic weather models were employed to generate daily precipitation time series for “ungauged” locations (methodology described in

| Component | Eigenvalue | % of variance | Cumulative % |
|-----------|------------|---------------|--------------|
| PC1       | 50.45      | 63.07         | 63.07        |
| PC2       | 4.79       | 5.99          | 69.05        |
| PC3       | 3.17       | 3.96          | 73.02        |
| PC4       | 2.25       | 2.82          | 75.84        |
| PC5       | 1.37       | 1.72          | 77.55        |
section 3.5). Figures 8 and 9 show the comparison results using two metrics (bias and RMSE) for annual total amount of precipitation, annual total number of wet-days, maximum consecutive wet and dry days. Each boxplot shows the range in statistical results from all stations. Bias values closer to zero and lower RMSE values imply that a model more accurately simulates the precipitation time series. Hence, the results shown in these figures suggest that Model-3 provides the most accurate simulation of annual precipitation amounts and occurrences. Because the average annual total amounts over the region lie between 1,000 mm and 1,500 mm, the error of Model-3 is less than 10%.

Figure 9 demonstrates that Model-3 is also superior for simulating sequences of precipitation occurrences (e.g., dry-spell and wet-spell). The figure demonstrates that Model-3 has the smallest biases and smallest RMSEs for both the annual mean number of wet days, and for the maximum number of consecutive wet days.

4.4 | Performance of the ungauged precipitation model

Table 5 shows the numerical evaluation results using three metrics (FC, PWC, and CSI) for validation at the Grahamsville station (methodology described in section 3.6). The high values of FC and CSI are 87.3 and 64.5%, and the low values are 83.9 and 55.7%, respectively. Harpham and Wilby (2005) also employed these metrics to assess their model performances. In their applications, the values of FCs are the range between 70 and 82%, and those of CSI are between 30 and 50%. The proposed stochastic model in this study yielded higher values of the evaluation metrics than the
**FIGURE 5** Catskill Mountains region delineated by PCA (based on precipitation amounts) for homogeneous daily precipitation with the period 1949–1959

**FIGURE 6** Regions delineated by OFA performed independently on each PCA region shown in Figure 6 for hydrologically homogeneous daily precipitation with the period 1949–1959 in the Catskill Mountains region.
analyses conducted by Harpham and Wilby (2005). This result implies that the proposed approach is able to provide an accurate description of the precipitation occurrences for this station as indicated by the high values of the evaluation metrics.

Figure 10 shows observed and simulated time series of annual and monthly means of precipitation amount and number of wet-days during the calibration period. In each panel the solid line represents the observed values, blue dotted line denotes the median values of the 20 ensembles, and the range (shaded blue) represents the upper and lower limits of the simulation. In all panels the observed time series falls within the simulated range for most data points, and the model captures the interannual variability, interannual trends, and the shape of the annual cycles of both precipitation amount and number of wet days.

Figure 11 shows results similar to Figure 10 except for the validation period. Note that the stochastic model used in this figure for the validation period uses parameters estimated during the calibration period only: no data from the validation period was used to calibrate the parameters. The graphical results show on average good agreement between
observed statistics and the medians of 20 ensembles over both calibration and validation periods. More specifically, the values for the calibration are closer to the observed values. It would be because the weather model for the calibration period uses more information from more stations available than for the validation period. These results suggest that the proposed stochastic weather model for ungauged daily precipitation sequences is able to describe accurately the annual and monthly statistics of observed data during periods outside of the validation period.

### 4.5 Climatic differences between regions

Figure 12 shows monthly mean precipitation amount and monthly mean number of wet-days during the calibration period for each region. The results shown in this section include statistics for each region calculated from observations, not from the stochastic model. The purpose of showing these results is to demonstrate that the regionalization identified by the model is indeed able to separate regions with different statistical characteristics.

OR1 (near the Pepacton Basin), OR2 (near the Ashokan basin and the southeastern section of our study area), and OR3 (near the Rondout and Neversink Basins) receive more precipitation than other regions during all seasons except summer. However, these do not exactly correspond to the region with the most frequent wet days: the most wet-days are observed in OR1 (the Pepacton Basin), OR4 (the Canonsville Basin), and OR5 (the northern section of our study area).

In the lower elevations of Ashokan and Rondout (OR2), winter through spring mean precipitation is higher than other regions while the number of wet-days is lower than other regions, indicating more frequent high intensity precipitation events. In general, western regions (OR1, OR4, and OR5) have more precipitation days than eastern regions (OR6, OR8, and OR11). Two sub-regions (OR6 and OR8) of the Schoharie watershed experience similar precipitation amounts, but differ in precipitation occurrence. Here, OR6 includes the higher elevation northeastern escarpment of the Catskill Mountains, where larger orographically moderated storms tend to occur.

The regional differences described above result from the spatial distribution of large precipitation events associated with coastal storms that mostly affect the southern and southeastern portions of our region; and the more frequent

### Table 5

Accuracy of the precipitation occurrence model for Grahamsville in Region 3 based on annual fraction correct (FC), annual percentage of wet-day correct (PWC), and annual critical success index (CSI) for both the calibration (1950s) and validation (1980s) periods, respectively.

|          | FC (%) | PWC (%) | CSI (%) |
|----------|--------|---------|---------|
| Calibration | 87.3   | 72.9    | 64.5    |
| Validation | 83.9   | 67.4    | 55.7    |
events associated with mid-latitude systems moving coming from the west or northwest, sometimes picking up moisture over the Great Lakes, that predominantly affect the western and northwestern portions of the study area; as documented descriptively by Thaler (1996) and analytically by Towey et al. (2018).

5 | CONCLUSION

In this study, a novel statistical approach based on the combined application of PCA and OFA is used to identify regions of “homogeneous” precipitation, in the sense that stations within regions have similar statistics of precipitation occurrence and amount. PCA is employed to identify stations with correlated precipitation amount, the results of which allow identification of regional patterns. PCA-based regions are referred to as “groups” to distinguish them from the OFA-based regions. OFA is then applied within each PCA-based group, providing further regionalization at finer spatial scales based on precipitation occurrence.

The method was calibrated using daily precipitation data from a network of 80 gauged stations in Catskill Mountains region for the period 1949–1959. This period was selected

FIGURE 10 Observed and estimated annual/monthly means of precipitations and the number of wet-days for the calibration period (1949–1959) at the selected station (Grahamsville). The range denotes the maximum and minimum values of the 20 ensemble time series, blue dotted line does the median values of the 20 ensembles, and the black solid line represents the observed values. (a) and (b) are annual time scale, and (c) and (d) monthly time scale.
In order to maximize the number and density of active stations, overall, 11 regions within which precipitation variations have coherent variations are delineated. Validation results from a more recent period (1981–1991), with fewer stations, confirms that the regions have not changed over recent decades. These results suggest that physical geography plays an important role in demarcating not only hydrological regimes (watersheds) but also regions with distinct climatic characteristics. Differences in average number of wet-days in the delineated regions are more marked than differences in monthly average precipitation amount.

Furthermore, these regions closely match the authors’ understanding of spatial climatic variations in this region based on experience working with data as well as published literature. These results are consistent with the hypothesis that storm tracks and orographic effects in combination are the primary factors in determining the spatial precipitation pattern for storms.

We validate the regionalization approach based on its accuracy in simulating daily precipitation series at ungauged locations. After eliminating one station for which data was available during both the calibration (1950s) and validation

**FIGURE 11** Observed and estimated annual/monthly means of precipitations and the number of wet-days for the validation period (1981–1991) at the selected station (Grahamsville). The range denotes the maximum and minimum values of the 20 ensemble time series, red dotted line does the median values of the 20 ensembles, and the black solid line represents the observed values. (a) and (b) are annual time scale, and (c) and (d) monthly time scale.
(1980s) periods, we estimated daily precipitation series using the proposed stochastic model at the location of the eliminated station. It was found that the model successfully estimates precipitation characteristics, including precipitation occurrence as well as precipitation amount, during both periods. Moreover, a comparative analysis indicates that the model including both PCA and OFA provides statistically significant improvement in the estimates of daily precipitation over the model with PCA only.

An important application of regionalization is the transfer of hydrological information from one station to another, or to a region where data are absent. Conventional approaches, based only on precipitation amount, do not consider spatial variation in the wet/dry-day at a specific location. The approach presented here addresses this limitation by including precipitation occurrence in addition to amount. These results demonstrate that this multivariate statistical approach provides information on the spatial patterns of precipitation that is unavailable from precipitation amount alone. They also demonstrate that a stochastic precipitation model can be used in conjunction with the multivariate methods to estimate the daily precipitation series at an ungauged site. Moreover, the application results of the current study region (Catskill Mountain Region) and the previous study region (Korean Peninsula) in Yeo (2014) have indicated that the proposed stochastic precipitation model could be useful for generating unmeasured daily precipitation series.

Despite the importance given to the accurate information for watersheds situated in mountainous regions, the application of SWGs or weather generators for mountainous regions are relatively uncommon because the increased heterogeneity of climatic conditions in geographically complicated areas is likely to diminish model accuracy (Breinl et al., 2017). Acharya et al. (2017) applied a SWG (WeaGET) (Chen et al., 2012) to the same study region without considering a spatial component. The proposed model in this study generates daily information based on observed daily precipitation at nearby stations, thus it is possible to apply to other regions.

While most SWGs such as LARS-WG (Semenov and Barrow 2002), WGEN (Richardson and Wright, 1984), and WeaGET (Chen et al., 2012) synthesize daily hydrometeorologic information using parameters of a selected distribution and a transition probability matrix with the Markov Chain, the proposed model generates daily information based on observed daily precipitation at nearby stations resulting in improved simulations of both precipitation amount and occurrence. However, we acknowledge that the accuracy of
this model may be sensitive to the homogeneity of the region and the number of observations within the region. In subsequent work, we will further refine our stochastic model by including geographical information in the model, and evaluate the applicability of this model to provide improved estimates of precipitation, and improved modelling of streamflow, over basins of varying spatial scales, from ungauged sub-basins with areas on the order of 10 km² to the entire watershed region.

ORCID

Myeong-Ho Yeo © https://orcid.org/0000-0002-6488-5309
Allan Frei © https://orcid.org/0000-0001-6295-5708

REFERENCES

Acharya, N., Frei, A., Chen, J., DeCristofaro, L. and Owens, E.M. (2017) Evaluating stochastic precipitation generators for climate change impact studies of New York City's primary water supply. Journal of Hydrometeorology, 18(3), 879–896.

Asong, Z., Khalili, M. and Wheeler, H. (2015) Regionalization of precipitation characteristics in the Canadian Prairie Provinces using large-scale atmospheric covariates and geophysical attributes. Stochastic Environmental Research and Risk Assessment, 29(3), 875–892.

Asong, Z.E., Khalili, M. and Wheeler, H. (2016) Multisite multivariate modeling of daily precipitation and temperature in the Canadian Prairie Provinces using generalized linear models. Climate Dynamics, 47(9–10), 2901–2921.

Baeriswyl, P.A. and Rebetez, M. (1997) Regionalization of precipitation in Switzerland by means of principal component analysis. Theoretical and Applied Climatology, 58(1), 31–41. https://doi.org/10.1007/BF00867430.

Bartholomew, D.J., Knott, M. and Moustaki, I. (2011) Latent Variable Models and Factor Analysis: A Unified Approach, Vol. 904. Chichester, West Sussex: John Wiley & Sons.

Breinl, K., Di Baldassarre, G., Lopez, M.G., Hagenlocher, M., Vico, G. and Rutgersson, A. (2017) Can weather generation capture precipitation patterns across different climates, spatial scales and under data scarcity? Scientific Reports, 7(1), 5449.

Burns, J.I. (1953) Small-scale topographic effects on precipitation distribution in San Dimas experimental Forest. Eos, Transactions American Geophysical Union, 34(5), 761–768.

Chen, J., Brissette, F. and Leconte, R. (2012) WeaGETS—a Matlab-based daily scale weather generator for generating precipitation and temperature. Procedia Environmental Sciences, 13, 2222–2235.

Comrie, A.C. and Glenn, E.C. (1998) Principal components-based regionalization of precipitation regimes across the Southwest United States and northern Mexico, with an application to monsoon precipitation variability. Climate Research, 10(3), 201–215.

Daly, C., Halbleib, M., Smith, J.J., Gibson, W.P., Doggett, M.K., Taylor, G.H., Curtis, J. and Pasteris, P.P. (2008) Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. International Journal of Climatology: A Journal of the Royal Meteorological Society, 28(15), 2031–2064.
fuzzy clustering approach. *Journal of Hydrology*, 405(3–4), 462–473. https://doi.org/10.1016/j.jhydrol.2011.05.044.

Schuurmans, J. and Bierkens, M. (2006) Effect of spatial distribution of daily rainfall on interior catchment response of a distributed hydrological model. *Hydrology and Earth System Sciences Discussions*, 3(4), 2175–2208.

Semenov, M.A. and Barrow, E.M. (2002) *LARS-WG: A Stochastic Weather Generator for Use in Climate Impact Studies Version 3.0, user manual*.

Sivapalan, M. (2003) Prediction in ungauged basins: a grand challenge for theoretical hydrology. *Hydrological Processes*, 17(15), 3163–3170. https://doi.org/10.1002/hyp.5155.

Thaler, J.S. (1996) *Catskill Weather*. Fleischmanns, NY: Purple Mountain Press Ltd.

Towey, K.L., Booth, J.F., Frei, A. and Sinclair, M.R. (2018) Track and circulation analysis of tropical and extratropical cyclones that cause strong precipitation and streamflow events in the New York City watershed. *Journal of Hydrometeorology*, 19(2018), 1027–1042. https://doi.org/10.1175/JHM-D-17-0199.

White, D., Richman, M. and Yarnal, B. (1991) Climate regionalization and rotation of principal components. *International Journal of Climatology*, 11(1), 1–25.

Wilby, R., Greenfield, B. and Glenny, C. (1994) A coupled synoptic-hydrological model for climate change impact assessment. *Journal of Hydrology*, 153(1–4), 265–290.

Wilby, R., Tomlinson, O. and Dawson, C. (2003) Multi-site simulation of precipitation by conditional resampling. *Climate Research*, 23 (3), 183–194.

Wilks, D.S. (1998) Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210(1–4), 178–191. https://doi.org/10.1016/s0022-1694(98)00186-3.

Wilks, D.S. (2011) *Statistical Methods in the Atmospheric Sciences*, 3rd edition. Oxford: Waltham, MA: Academic Press.

Wilks, D.S. and Wilby, R.L. (1999) The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, 23(3), 329–357. https://doi.org/10.1177/030913339902300302.

Yeo, M.-H. (2014) *Statistical modeling of precipitation processes for gaged and ungaged sites in the context of climate change*. PhD Thesis, McGill University, Montreal.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Yeo M-H, Frei A, Gelda RK, Owens EM. A stochastic weather model for generating daily precipitation series at ungauged locations in the Catskill Mountain region of New York state. *Int J Climatol*. 2020;40:687–705. [https://doi.org/10.1002/joc.6230](https://doi.org/10.1002/joc.6230)