Simulating Precipitation in the Northeast United States Using a Climate-Informed K-Nearest Neighbors Algorithm

Saman Armal[1,2,3], **Naresh Devineni** [1], **Nir Y. Krakauer** [1,2,3], **Reza Khanbilvardi** [1,2,3]

[1] Department of Civil Engineering, The City University of New York (City College), New York, NY 10031

[2] NOAA-Cooperative Center for Earth System Sciences and Remote Sensing Technologies, The City University of New York (City College), New York, NY 10031

[3] Center for Water Resources and Environmental Research, The City University of New York (City College), New York, NY 10031

KEYWORDS: Non-stationarity, Climate-informed K-NN model, Precipitation

Abstract

Decadal prediction using climate models faces long-standing challenges. While global climate models may reproduce long-term shifts in climate due to external forcing, in the near-term they often fail to accurately simulate interannual climate variability, as well as seasonal variability, wet and dry spells, and persistence, which are essential for water resources management. We developed a new climate-informed K-Nearest Neighbor (K-NN) based stochastic modeling approach to capture the long-term trend and variability while replicating intra-annual statistics. The climate-informed K-NN stochastic model utilizes historical data along with climate state information to provide improved simulations of weather for near-term regional projections. Daily precipitation and temperature simulations are based on analog weather days that belong to years similar to current year's climate state. The climate-informed K-NN stochastic model is tested using 53 weather stations in the Northeast United States with an evident monotonic trend in annual precipitation. The model is also compared to the original K-NN weather generator and ISIMIP-2b GFDL General Circulation Model in a cross-validation mode. Results indicate that the climate-informed K-NN model provides improved simulations for the dry and wet regimes, and better uncertainty bounds for annual average precipitation. The model also replicates the within-year rainfall statistics. For the 1960-1970 dry regime, and the model captures annual average precipitation and the intra-annual coefficient of variation. For the 2005-2014 wet regime, the model replicates the monotonic trend and daily persistence in precipitation. These improved simulations can be used for accurately simulating near-term streamflow, which in turn can be used for short-term water resources planning and management.

1. Introduction

Stationarity of a hydroclimate time series is defined as the invariance of its statistics

with time (Shaw, 2014). Any time variation in the properties can indicate non-stationarity, including shifts in the mean (Westra et al., 2013; Alexander et al., 2006; Fischer & Knutti, 2014; Villarini et al., 2009; Aziz & Burn, 2006), variance (Coulibaly & Burn, 2004; Lewis & King, 2017), and the autocorrelation structure (Razavi et al., 2015). In their much-acknowledged work, Milly and co-authors (Milly et al., 2008) urge the water resources community to reconsider stationarity as a central assumption of risk assessment and planning analysis. They suggest that in the wake of substantial anthropogenic change of Earth's climate, stationarity is dead as a viable assumption. However, Jain and Lall (Jain & Lall, 2001) argue that the presence of quasi-cyclical modes of natural climate variability render the traditional assumption of stationarity void in any case. Essentially, even in a stationary climate, one might observe statistically significant trends in hydro-climatic systems over time due to natural variability (National Research Council, 1999; Cohn & Lins, 2005) or processes with long-term persistence (Villarini et al., 2009). Also, given short observation periods, part of a long-timescale oscillation can be wrongly extrapolated as a monotonic trend (Bloschl & Motanari, 2010). Lin and co-authors (Lins & Cohn, 2011) also show that non-stationarity is not always discriminable from stationarity. They argue that this question is highly dependent on which signals are sampled and the length of the period under investigation.

Many researchers have adopted the use of General Circulation Models (GCMs) as a means for the planning and risk assessment of hydro-systems under changing and uncertain future climates. Scenarios of future climate are often based on long-term GCM simulations forced by different emission pathways, with bias correction for systematic errors based on historic data (Wilks & Wilby, 1999; Steinschneider et al., 2015; Taner et al., 2017; Trzaska & Schnarr, 2014). However, even when GCMs correctly represent the long-term secular trend, they may fail to capture internal variability (Frederick, 2011; Hempel et al., 2013, Cassou et al., 2018) and simulating extremes (Katiraie-Boroujerdy et al., 2019). This can lead to substantial bias in representing climate, particularly for near-term projections at the regional scale (Van Oldenborgh et al., 2013; Krakauer & Fekete, 2014). Moreover, bias correction does not provide any reliable solution to fix this disparity because it assumes that the bias statistics calculated over the historical period can be extrapolated to the future (Kerkhoff et al., 2014). Given the uncertainties involved in GCMs, the question arises to what extent their application is reliable in different regions. Are GCM simulations sufficient for future hydro-systems planning and management? Or, should we rely on the assumption of stationarity for robust decision-making, until current models are improved

(Stakhiv, 2011)? We explored answers to these questions using the Northeast United States climate region as a case study. The Northeast climate region has significant observed trends in climate and is one of the wettest parts of the United States. Over the last century, long-term trends in precipitation have been 9.5 mm ± 2 mm/decade, mainly in the spring, summer and fall seasons (Hayhoe et al., 2007). Substantial upward trends are also noted in extreme precipitation, based on the recent analyses of the Northeast climate (Hoerling et al., 2016; Easterling et al., 2017; Huang et al., 2017). This increase of the total amount of precipitation and frequency of heavy precipitation events raises concerns about flooding and its effects on aging infrastructure in the Northeast (Horton et al., 2014).

Figure 1 summarizes different statistics of daily precipitation for a bias-corrected historical GCM simulation (here GFDL model from ISIMIP2-b dataset (Warszawski et al., 2014; Frieler et al., 2017)) and contrasts them with in-situ observations in 53 selected weather stations over the Northeast climate region (the in-situ data are described in the next section: Data and Methodology). The climate model simulations are obtained from the grid cells corresponding to the 53 weather stations to enable comparison across datasets. The median of average precipitation in the model simulation is much higher than that of the corresponding stations. This deviation is more prominent in the 1960s, although, it has to be noted that the above climate model is not forced with observed sea surface temperatures (SSTs), hence, accurate representation of the time properties may not fully be possible. The range of the observations, representing the across station variation for each year, is greater than that found in the bias-corrected climate model, indicating that the GCM simulations fail to replicate the spatial variability. As well, Figure 1-c (comparison of the intra-annual coefficient of variation (CV)) confirms that the observed intra-annual (i.e., within the year) variability is not well captured in the model. Other statistics including annual skewness, trends in average annual rainfall (using Mann-Kendall Tau values), and annual lag-1 auto-correlation also show significant biases. The model simulates the length of the dry spell relatively well but cannot capture the extremes (exceeding 95 percentile of non-zero daily precipitation computed for each dataset independently) and wet spell length. These limitations were seen in other bias-adjusted GCMs as well, including HadGEM2-ES, IPSL-CM5A-LR, and MIROC5 (see Supplementary Material). Owing to these shortcomings, GCMs are less effective in their application for water resources planning and management. Here, we suggest that the utilization of long historical data in conjunction with climate state information provides a more reliable tool to simulate daily weather variables in the near term, and our study explores the evidence for this assertion by applying

a climate-informed statistical Weather Generator (WG) to simulate precipitation over the Northeast.

WGs are intended to produce synthetic weather sequences that mimic the statistical properties of observed meteorological records (Wilks & Wilby, 1999). Different parametric and semi-parametric WGs are available. WGs may use an autoregressive modeling framework (Aiyesimoju, 2010) or pre-clustering of rainfall cells/points to simulate the storm arrival in a generalized linear process (Onof et al., 2000; Mannshardt-Shamseldin et al., 2010). They may also employ a hierarchical framework, which conditions the local meteorology on large-scale synoptic climatological patterns and weather types or regimes (Allard et al., 2015; Benestad, 2016; Pierce et al., 2014). Non-parametric methods for WGs often use resampling techniques to simulate synthetic data from the observations (Oriani et al., 2014; Pierce et al., 2014; Yiou, 2014). A well-established method among the non-parametric methods for WGs is the K-Nearest Neighbors (K-NN) method. The idea of K-NN can be traced back to the concept of "discriminant space" developed originally by Young (Young, 1994). He used an orthogonalized multi-dimension space as a predictor to choose the past days that most resembled the current weather condition. Similarly, Lall and Sharma applied discrete kernel weighing to select the nearest neighbors in the historical data (Lall & Sharma, 1996; Sharma et al., 1997). Their work was extended by the inclusion of a large set of weather variables (Rajagopalan & Lall, 1999); modified by using Mahalanobis distance in the weighing function (Yates et al., 2003); and improved by considering inter-station correlation (Sharif & Burn, 2007; King et al., 2015). However, WGs commonly rely on the assumption of stationarity in weather generation processes and therefore, cannot capture shifts in the statistics of hydrologic variables (Benoit et al., 2018). A few studies addressed this limitation by adding simulated standardized anomalies to cyclo-stationary mean (Smith et al., 2017), or by incorporating non-stationary weather generation parameters (Chandler, 2005; Lima et al., 2015).

In this study, we develop and present a new climate-informed K-NN algorithm to simulate future weather. The methodology we employ is exploratory, where we examine and incorporate the assumption of non-stationarity with different training and validation periods. Moreover, we evaluate and compare the skill of the proposed climate-informed resampling scheme with the outputs from the original K-NN method and with the ISIMIP-GFDL model outputs. We look at the ability of the model to simulate the 1960s drought and the recent wet climate in the Northeast USA and argue that utilization of historical data along with information from climate may more reliably replicate both secular trends and

internal variability in the data for short-term hydrologic planning purposes.

2. Data and Methods

Our recent analysis of extreme rainfall in the United States characterizes the Northeast region as one with a significant shift in the frequency of extreme rainfall events (Armal et al., 2018). The meteorological data used in this study is taken from 53 stations with an identified monotonic trend in this analysis. This dataset was derived from the Global Historical Climatology Network - Daily (GHCN-D) database (https://www.ncdc.noaa.gov/oa/climate/ghcn-daily/). The 53 selected stations have at least 92 years of complete precipitation data during 1900 - 2014 (115 years). The fraction of missing data is less than 20%. Figure 2 presents the spatial distribution of these stations. The number of years of available data is shown using the color bar, and the monotonic trend in the annual total precipitation measured using Mann-Kendall tau is displayed using the size of the circle. For applying our algorithm, we also acquired the maximum and minimum daily temperature (Tmax and Tmin) from the same stations. Since the filtering process was only based on having a trend in precipitation, the daily temperature data for many stations were not fully available. As the focus of the study is to analyze the daily precipitation, in our simulation, we used the spatial average of the maximum and minimum temperature as the Tmax and Tmin data for all the stations. In other words, we adopt a single time series (averaged over the 53 stations) for daily Tmax and Tmin. Since the temperature is spatially homogeneous, we believe this is a reasonable compromise due to the lack of data.

The proposed Climate Informed K-NN resampling model is an extension of the K-NN weather generator. In the next sections, we briefly describe the original K-NN algorithm and then introduce the proposed climate-informed resampling scheme.

2.1. The K-NN Weather Generator

The K-NN weather generator (originally developed by Rajagopalan and Lall (Rajagopalan & Lall, 1999)) is a data-driven approach that simulates future weather variables conditional on the current weather state and its relation to historical weather. It has the following steps.

1. The feature vector comprising the current weather variables is first defined. We call it the conditioning vector. In our case, the feature vector or the current conditioning vector is $V_i = [P_i, Tmax_i, Tmin_i]$ for each weather station, for the current day i.

2. For this conditioning vector (V_i), we compute its distance, in the state space, to the historical weather vectors using the Mahalanobis distance metric.

$$d_{ij} = \sqrt{(V_i - V_j)^T \Sigma^{-1} (V_i - V_j)}$$

Σ is the covariance matrix of weather variables in the corresponding season. V_j is the historical weather vector for a day j

3. The historical weather vectors $V_j s$ are ordered/ranked according to the distance d_{ij}, and the K nearest neighbors are identified. Weather vectors that have smaller (larger) distances indicate similarity (difference) in terms of the weather conditions. For each of these K nearest neighbors, we identify the successor vector that comprises the next day's values of the weather.

4. A discrete kernel probability (Lall & Sharma, 1996) is defined for each of the K neighbors using the following function.

$$K[j(i)] = \frac{(1/j)}{\sum_{j=1}^{k}(1/j)}$$

$K[j(i)]$ is the probability with which V_j is resampled for the current day i. Closer neighbors have more probability of being resampled. This resampling kernel is the same for any day i.

5. As the final step, one of these vectors is resampled according to the kernel probability, and its successor vector is taken as the weather for the next day.

6. With the immediately generated weather as the current feature vector, this process is then repeated to simulate the weather for the following day.

7. We prescribed $K = 45$ since that choice gave us the lowest absolute error in simulating annual precipitation. Generally, it is suggested, based on the asymptotic argument, that K should be proportional to $n^{d/(d+4)}$, n being the total number of neighbor vectors in the space, and d being the dimension of the feature vector (Fukunaga, 1990). We investigated the sensitivity of our simulation results to number of neighbors, and found that number of neighbors did not considerably alter the absolute error. The optimal neighbors we used are mostly conforming to the ones recommended in the literature. We also

observed a similar low error across all the neighbors and windows used, indicating that the error is uniform across all windows

2.2. Climate-Informed K-NN Resampling

As presented in section 2.1, the original K-NN scheme is based on the nearest neighbors of the historical daily weather variables. We modify this scheme by incorporating large-scale meteorological information in the choice of neighbors and the resampling process using the premise that this large-scale meteorological information drives the variability (secular and cyclical). We use two climate variables in the model: the annual average of 500-hPa geopotential height anomaly (GPH) and the North Atlantic Oscillation (NAO) index from 1900 - 2014. We extracted the long-term mean of GPH over the applied weather stations from NOAA-CIRES Twentieth Century Reanalysis-V2c dataset (provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA) available from NOAA/ESRL website (https://www.esrl.noaa.gov/psd/). This dataset is the result of assimilation of observations into a numerical weather prediction model with an Ensemble Kalman Filter (Compo et al., 2011). It is available on a global grid, with a spatial resolution of about 200 km. The NAO index, aggregated from monthly to yearly averages, is obtained from NOAA Climate Prediction Center (http://www.cpc.ncep.noaa.gov/data/teledoc/nao.shtml). Studies confirm the potential of applying NAO to predict eastern United States' climate and its association with the 1960s drought and the following wet period (Seager et al., 2012; Bradbury et al., 2002; Coleman & Budikova, 2013). Our recent studies also demonstrated that variation in NAO partly explains the frequency of precipitation extremes in the Northeast (Armal et al., 2018, Armal and Khanbilvardi, 2019).

In the vector of climate variables ($C = (NAO, GPH)$), Mahalanobis distance provides a metric to identify climate similarity. We now have two distance measures: one is identifying weather similarity (d) from the days, and one is identifying climate similarity (d_c) from the years. We use these two in conjunction to reorder the weather neighbors in accordance with their climate similarity. Weather neighbors that belong to a climate similar to the current year's climate will be ranked ahead of the weather neighbors that are further away from the current climate. The kernel density function and resampling is now applied to the reordered neighbors to choose the weather neighbors that belong to the most analogous years (i.e., most similar

climatic years).

Figure 3a depicts the main steps of the climate-informed model. The model consists of three key steps, which are explained here using an illustration shown in Figure 3b. (1) The algorithm identifies the current state and applies Mahalanobis distance to obtain the closest members of historical weather vector to the current condition and stores their time-indices ($x_{i,m}$). In the illustration, we show 14 nearest neighbors for the current weather variable $[2.0, 28.8, 14.2]$ under the R, $Tmax$ and $Tmin$ columns. (2) The algorithm applies Mahalanobis distance on yearly climate vector to find the most analogous climate years (C_m). The closest members (to the current condition) from the historical weather vector that are also close in terms of the similar climate index values are then prioritized. From the illustration, the current climate vector is $[GPH, NAO] = [0.91, 0.71]$. This is in the year 1982, and the closest neighbors to this climate state are years 1946, 1920, 1953, 1914, etc. These rows are highlighted to show climatically similar weather vectors. They are prioritized, i.e., they move up in order as shown in the table on the right. Notice how the 12th neighbor by weather distance d now becomes the 3th closest neighbor after re-ordering by the climate distance (dc). There is now a greater probability that this neighbor will be sampled. (3). The algorithm applies a discrete kernel estimator $K[j(i)]$ on the relative frequency of the data lying in the local neighborhood and resamples one of $x_{i,m}$ from climate-rearranged set of nearest neighbors. The successor vector of the selected neighbor is used as the weather of the next step.

2.3. Model Training

Similar to other memory-based algorithms, providing more training data to the K-NN model reduces the chance of misclassification in the selection of neighbors (Friedman et al., 2001). As the size of training data increases, there are examples to generalize to the unknown sample and generate a good local approximation. If the properties of the current state and near-term projections are similar to the properties of the training data (historical data in this case), the algorithm yields a more reliable result. By contrast, when the process is not stationary, and the simulation period is not represented in the training dataset, the model predicts different statistics than observed. We experiment with this idea and evaluate the model's performance using a range of out-of-sample analysis windows.

We train both the original and climate-informed K-NN resampling models over

expanding window sizes that begin in 1900 and end in years ranging from 1940 to 2004. For each training window, the model simulates daily weather for the succeeding ten years; i.e., we simulate weather data from 1941 to 1950 using the historical data from 1900 to 1940 as the training set; likewise, we simulate weather data from 1942 to 1951 using the historical data from 1900 to 1941 as the training set. We do this in moving windows that end with generating weather data for 2005 to 2015 using training data from 1900 to 2004. As we move from 1941 to 2005, the size of the historical data (training set) increases one year at a time. We repeat this exercise 30 times to get an ensemble of synthetic weather. Hence, in a series of simulations, we increase the size of training data and consequently evaluate the skill of the model in generating precipitation. The step-wise training provides a systematic inspection of the power of the algorithm over different segments of data, with and without climate conditioning. All the variables were first de-seasoned by removing the calendar day's mean. In the next section, we discuss the results by pooling the outputs from all the weather stations.

3. Results

Figure 4 shows the annual distribution of average precipitation as simulated using the step-wise training approach. The shaded area in Figure-4(a) and Figure-4(b) indicates the range of decadal simulation runs across all the stations from the original K-NN and the climate-informed K-NN models, respectively. These boundaries are smoothed over the decadal periods using a locally weighted smoothing approach (LOWESS) (Loader, 1999). The red and the blue lines represent the LOWESS applied to the median of the annual distribution. The boxplots represent the annual average observed precipitation over each of the 10-year simulated blocks.

For the original K-NN scheme (Figure 4-a), the decadal simulations exhibit a positive bias and do not accurately represent rainfall deficits, especially in the periods up to the early 1970s. The increase in the size of the training data certainly reduces the bias, as seen in the simulations of recent decades. A larger training size improves the choice of neighbors and the generation of the annual average daily precipitation. Nevertheless, K-NN simulation depicts a relatively high bias around and after the 1960s. The shift in the characteristic rainfall distribution due to a severe drought that occurred in the early to mid-1960s explains this bias. This abrupt drought was followed by a wetter climate in the region that began around the early 1970s and

has continued since (Seager et al., 2012). The simulation blocks that start in the wet period after the 1960s show a more reliable outcome.

Figure 4-b shows the outcome of the climate-informed model. Adding climate information generates a larger spread of average annual daily precipitation and broadens the envelope of the range of values. Compared to the original K-NN algorithm, the climate-informed K-NN model covers the observed data in low precipitation events and simulates the 1960s annual precipitation well. The findings of Seager et al. (Seager et al., 2012) suggest that the temporary drought is a result of the oscillatory nature of NAO and the seesaw in the pressure and GPH anomalies between the subpolar and subtropical Atlantic Oceans. Improved simulations of the 1960s drought decade illustrate that including these components of the climate state is beneficial in projecting dry (or wet) decades reliably. The climate-informed model makes use of analogous years, in this case, ones with dry conditions that are available in the training dataset but are not necessarily used in the K-NN resampling. However, it is apparent from Figure 4-b that the drying condition in climate-informed simulation continues longer than observation – until circa 1970. At each step of training, the window expands one year, and the updated historical data overlaps with low precipitation years in the 1960s. This will affect the simulation and prolong the dry condition in the outcome for a few more simulation windows.

To evaluate the performance of the developed models, we compare the skill of the original and the climate-informed K-NN scheme over a range of statistics that may be of interest to water managers. These properties are calculated for 10-yr periods, succeeding two different training periods: (1) The training period ending in 1960 (1900-1960), to address the ability of the models to simulate the 1960s drought (1961-1970), and (2) The entire period of the dataset, excluding the last 10 years of data (1900-2004), to simulate the continued contemporary wet period (2005-2014). Figure 5 shows the outcome of 1961-1970 precipitation simulations using 1900-1960 as the training data. Figure 6 compares the same statistics for the 2005-2014 simulation period that uses 1900-2004 for training. The variation in these statistics across the 30 iterations and the 53 stations is shown in the boxplots. Outliers are excluded. The statistics are obtained yearly from daily precipitation and include the average intra-annual coefficient of variation, skewness, trends in annual average precipitation (Mann-Kendall Tau), extreme values above the 95th percentile, wet spell and dry spell length with 1-mm threshold, and lag-1 autocorrelation. The application

of K-NN resampling previously showed its ability to reproduce such statistics over historical simulations (Rajagopalan & Lall, 1999). Here it's worth mentioning that GCM runs are driven by radiative forcing such as greenhouse gas concentrations and volcanic eruptions and are not expected to capture the timing of decadal climate variability that is largely internal to the atmosphere/ocean (Fernandes et al. 2015, Wang et al. 2017, Trenberth et al. 2018). Furthermore, the current work does not intend to assess the ability of the climate models based on a set of time-dependent statistics. Rather, it aims to study the skill of the proposed model using the output of the trend preserving bias-correction methods (Hempel et al. 2013) as an additional baseline.

The GFDL statistics for the period 2005-2014 are obtained by combining the last year of the historical simulation (2005) with each available RCP pathways (2006-2014). Other statistics, including intra-annual CV and skewness, confirms the ability of the climate- informed scheme to reproduce 1960s data. The comparison of the model-simulated statistics is made with the observed statistics using a similarity measure with a bootstrap approach. From the distributions of the observed and the model-simulated statistics, we randomly draw 100 values and compare how many of them match, i.e., how many of the simulated statistics are equal to the observed statistics. We compute a similarity ratio as the fraction of common values in the bootstrap sample of 100. A ratio close to one (zero) indicates that most of the randomly sampled values from the model simulated, and the observed statistics are similar (different). We repeat this sampling and computing the similarity ratio 10,000 times to obtain a distribution of the similarity ratio. The 5th percentile from this distribution is selected as the test statistic and reported in Table 1 for both the 1960s and the recent wet decades. It is evident from the Table that the climate-informed K-NN model simulations compare better than the original K-NN and the GCM simulations in terms of the statistics used. For example, for the average annual precipitation, at the 95% confidence level, the match ratio is only 67.6% for K-NN and 59.7% for GCMs, but it is 94.5% for the climate-informed model, indicating that for the same chosen level of confidence, the climate-informed model has much higher match ratio – similarity to the observed. Trend compares poorly with the observations in both the decades across the three models. GCM similarity is generally lower compared to the two K-NN models. The largest improvement that the climate-informed K-NN model offers in terms of its similarity to the observed statistics is in the annual average precipitation during the 1960s decade. In the 2005-2014 decade, the most considerable improvement seems to be on the annual average precipitation and the intra-annual coefficient of variation of rainfall.

In addition to these comparisons, an examination of Figures 5 and 6 indicate that the climate informed model better reproduces the 1960s trend than the GFDL simulations. It is an example of the shortcoming of GCMs to capture the decadal fluctuations of annual precipitation due to internal variability in the climate system. The simulation of contemporary data also shows the capability of the climate-informed model in better simulating precipitation statistics, even though all the models poorly represent the recent trend. In the simulation of the 1960s, all three models (K-NN and climate-informed K-NN along with GFDL) capture the observed 95% precipitation extreme values. In the contemporary runs, the climate-informed scheme depicts a more reliable performance. Either with the short or long training period, both resampling schemes, as well as the GFDL model, successfully capture wet spell and dry spell length. In both sets of training, the original and modified K-NN schemes replicate the lag-1 auto-correlation. The GFDL results are also acceptable with the 1960s simulation, but not reliable in the contemporary period.

In Figure 5 and 6, we use different statistics to assess the performance of the climate-informed resampling, in two different observation blocks. In Figure 7 and 8, we measure the capability of the climate-informed model to simulate the decadal variability of the 1960s and the contemporary period (2005-2014). Comparing the inter-annual variability of average daily precipitation reveals that for the 1960s drought, the climate-informed model improves on the original K-NN resampling scheme, particularly with respect to annual average precipitation. For the contemporary period, both models generate a relatively similar pattern in the median of values, but only the climate-informed model preserves the range and tails of average daily precipitation well. Notably, K-NN underestimates the observed variability for both periods. The outcomes of the climate-informed model, on the other hand, relatively overestimate the values of the median.

In summary, as a result of incorporating climate information in the resampling model, we observed an improvement in the simulation of average annual precipitation over periods of changing regional hydroclimate regime. The use of climate data favorably modifies the resampling scheme and replicates several characteristics of daily precipitation, although limitations are observed in the preservation of tails in some statistics (e.g., wet spell and dry spell). Inclusion of additional regional scale atmospheric variables as climate informants may better resolve the tails in these statistics. Also, in the current model, due to lack of data, we applied the average temperature over the region, which can lead to a biased simulation that may be affecting the higher precipitation quantiles. The performance of the model may be more satisfactory if we incorporate historical temperature data which is

specific to each station, where available.

For both training periods, the climate-informed model regularly performs well in recreating different statistics over the periods of studies. Specifically, the model replicates the range of average annual daily precipitation values in both the 1960s and the contemporary wet periods. The Northeast is characterized by ongoing wetting shift (Huang et al., 2017) and an upward trend in extreme precipitation (Easterling et al., 2017). Krakauer (Krakauer, 2014) shows that the mean of the recent precipitation in parts of the Northeast is 25% above its value before 1970, with even larger increases in the intensity of wet extremes. The results of this simulation indicate that the long-size training dataset contains similar hydroclimatic years, which enable the simulation of these new prevailing conditions with climate-informed K-NN. The wet period observed in the earlier part of the twentieth century (Seager et al., 2012) conditions the algorithm in the later sets of simulations by supplying neighbors with analogous hydroclimate conditions.

4. Summary and Conclusion

Engineering design and practice may rely on the simple idea that the characteristics of future events resemble the past. Non-stationarity contradicts this traditional point of view. Practically, GCMs perform a critical role in addressing non-stationarity in meteorological data due to changing climate forcing. However, GCMs often fail to represent inter-annual variability accurately on decadal scale. Our analysis adopted an exploratory method to verify the feasibility of applying stochastic modeling in inferring these characteristics for precipitation in the Northeast USA.

We proposed a climate-informed K-NN model which adjusts the vector of selected nearest neighbors based on climatic information. We employed both the original K-NN and proposed climate-informed model and performed a set of simulations using a step-wise expanding training window, with a minimum length of 40 years. It is revealed that the climate-informed model replicates the range of annual average daily precipitation, while the original K-NN scheme fails to capture lower tail values across many blocks of simulation. With the simulation of different statistics of daily precipitation in the 1960s and contemporary (2005-2014) period, we compared the results of the proposed model with the original K-NN and the GFDL model. Results indicate that the climate-informed model consistently presents better performance. It demonstrates that incorporating climate information improved the skill of resampling scheme in the regeneration of different usable statistics.

Lins and co-authors (Lins & Cohn, 2011) argue that a lack of thorough understanding of the physical and scientific background in the context of hydrology blurs the line between stationarity and non-stationarity. They conclude that: "In a statistical sense, while the future will not repeat the past, its properties can be inferred from the past." In this study, we show that the utilization of historical data along with proven climate informants allows the simulation of the statistical properties of daily precipitation over the next ten years. The demand for such empirical correction that applies a data-adaptive scheme to improve near-term projections is suggested by (Krakauer, 2014; Krakauer & Fekete, 2014). Although GCMs may not always simulate these properties well over the near-term future, they may be a valuable tool to inform the predictors and initial condition for simulating future regional hydroclimate using climate-informed K-NN or similar methods.

Retrospective GCM runs, including re-analysis products, are potentially quite useful in assessing global teleconnections associated with regional hydroclimatology. Applying climate-informed resampling to regulate local neighbors with modes of natural climate variability relies on knowledge of modes of decadal variability of regional climate. This methodology offers a way to integrate climate information with historical variability for improved simulations. These improved simulations can be used for simulating near-term streamflow, which in turn are used for short-term water resources planning and management. Water managers and decision makers can benefit from this information for robust system design and water resources analysis.

Acknowledgments

This study is supported by:

1. National Oceanic and Atmospheric Administration Cooperative Center for Earth System Sciences and Remote Sensing Technologies (NOAA-CESSRST) under the Cooperative Agreement Grant #: NA16SEC4810008.
2. U.S. Department of Energy Early CAREER Award No.DE-SC0018124 for the second author (Naresh Devineni).

The statements contained within the article are not the opinions of the funding agency or the U.S. government but reflect the authors' views.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Aiyesimoju, K.O., 2010. An assessment of Box-Jenkins models: Forcados monthly rainfall as case study. *Journal of Applied Science, Engineering and Technology*, *10*.

Alexander, L.V., Zhang, X., Peterson, T.C., Caesar, J., Gleason, B., Klein Tank, A.M.G., Haylock, M., Collins, D., Trewin, B., Rahimzadeh, F. and Tagipour, A., 2006. Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research: Atmospheres*, *111*(D5).

Ailliot, P., Allard, D., Monbet, V. and Naveau, P., 2015. Stochastic weather generators: an overview of weather type models. *Journal de la Société Française de Statistique*, *156*(1), pp.101-113.

Armal, S., Devineni, N. and Khanbilvardi, R., 2018. Trends in extreme rainfall frequency in the contiguous United States: Attribution to climate change and climate variability modes. *Journal of Climate*, *31*(1), pp.369-385.

Armal, S. and Khanbilvardi, R., 2019. Anomalies in the US Precipitation Extremes and Their Association with Different Modes of Climate Variability. *Hydrological sciences journal*.

Aziz, O.I.A. and Burn, D.H., 2006. Trends and variability in the hydrological regime of the Mackenzie River Basin. *Journal of hydrology*, *319*(1-4), pp.282-294.

Benestad, Rasmus. "Downscaling climate information." *Oxford Research Encyclopedia of Climate Science*. 2016.

Benoit, L., Vrac, M. and Mariethoz, G., 2018. Dealing with non-stationarity in sub-daily stochastic rainfall models. *Hydrology and Earth System Sciences*, *22*(11), pp.5919-5933.

Blöschl, G. and Montanari, A., 2010. Climate change impacts—throwing the dice?. *Hydrological Processes: An International Journal*, *24*(3), pp.374-381.

Bradbury, J.A., Dingman, S.L. and Keim, B.D., 2002. New England drought and relations with large scale atmospheric circulation patterns 1. *JAWRA Journal of the American Water Resources Association*, *38*(5), pp.1287-1299.

Cassou, C., Kushnir, Y., Hawkins, E., Pirani, A., Kucharski, F., Kang, I.S. and Caltabiano, N., 2018. Decadal climate variability and predictability: challenges and opportunities. *Bulletin of the American Meteorological Society*, *99*(3), pp.479-490.

Chandler, R.E., 2005. On the use of generalized linear models for interpreting climate variability. *Environmetrics: The official journal of the International Environmetrics Society*, *16*(7), pp.699-715.

Cohn, T.A. and Lins, H.F., 2005. Nature's style: Naturally trendy. *Geophysical research letters*, *32*(23).

Coleman, J.S. and Budikova, D., 2013. Eastern US summer streamflow during extreme phases of the North Atlantic Oscillation. *Journal of Geophysical Research: Atmospheres*, *118*(10), pp.4181-4193.

Compo, G.P., Whitaker, J.S., Sardeshmukh, P.D., Matsui, N., Allan, R.J., Yin, X., Gleason, B.E., Vose, R.S., Rutledge, G., Bessemoulin, P. and Brönnimann, S., 2011. The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, *137*(654), pp.1-28.

Coulibaly, P. and Burn, D.H., 2004. Wavelet analysis of variability in annual Canadian streamflows. *Water Resources Research*, *40*(3).

Easterling, D.R., Kunkel, K.E., Arnold, J.R., Knutson, T., LeGrande, A.N., Leung, L.R., Vose, R.S., Waliser, D.E. and Wehner, M.F., 2017. Precipitation change in the United States.

Fernandes, Katia, Alessandra Giannini, Louis Verchot, Walter Baethgen, and Miguel Pinedo-Vasquez, 2015. Decadal covariability of Atlantic SSTs and western Amazon dry-season hydroclimate in observations and CMIP5 simulations. *Geophysical Research Letters* 42(16),

pp.6793-6801.

Fischer, E.M. and Knutti, R., 2014. Detection of spatially aggregated changes in temperature and precipitation extremes. *Geophysical Research Letters*, *41*(2), pp.547-554.

Frederick, K.D., 2011. Principles and concepts for water resources planning under climate uncertainty. *Journal of Contemporary Water Research and Education*, *112*(1), p.7.

Friedman, J., Hastie, T. and Tibshirani, R., 2001. *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

Frieler, K., Lange, S., Piontek, F., Reyer, C.P., Schewe, J., Warszawski, L., Zhao, F., Chini, L., Denvil, S., Emanuel, K. and Geiger, T., 2017. Assessing the impacts of 1.5 C global warming–simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b). *Geoscientific Model Development*.

Fukunaga, K., 1990. Introduction to statistical pattern recognition (2nd ed.). San Diego, CA, USA: Academic Press Professional, Inc.

Hayhoe, K., Wake, C.P., Huntington, T.G., Luo, L., Schwartz, M.D., Sheffield, J., Wood, E., Anderson, B., Bradbury, J., DeGaetano, A. and Troy, T.J., 2007. Past and future changes in climate and hydrological indicators in the US Northeast. *Climate Dynamics*, *28*(4), pp.381-407.

Hempel, S., Frieler, K., Warszawski, L., Schewe, J. and Piontek, F., 2013. A trend-preserving bias correction–the ISI-MIP approach. *Earth System Dynamics*, *4*(2), pp.219-236.

Hoerling, M., Eischeid, J., Perlwitz, J., Quan, X.W., Wolter, K. and Cheng, L., 2016. Characterizing recent trends in US heavy precipitation. *Journal of Climate*, *29*(7), pp.2313-2332.

Horton, R., Yohe, G., Easterling, W., Kates, R., Ruth, M., Sussman, E., Whelchel, A., Wolfe, D. and Lipschultz, F., 2014. Ch. 16: Northeast, climate change impacts in the United States. *The Third National Climate Assessment*, pp.371-395.

Huang, H., Winter, J.M., Osterberg, E.C., Horton, R.M. and Beckage, B., 2017. Total and extreme precipitation changes over the Northeastern United States. *Journal of Hydrometeorology*, *18*(6), pp.1783-1798.

Jain, S. and Lall, U., 2001. Floods in a changing climate: Does the past represent the future?. *Water Resources Research*, *37*(12), pp.3193-3205.

Katiraie-Boroujerdy, P.S., Akbari Asanjan, A., Chavoshian, A., Hsu, K.L. and Sorooshian, S., 2019. Assessment of seven CMIP5 model precipitation extremes over Iran based on a satellite-based climate data set. *International Journal of Climatology*, 39(8), pp.3505-3522.

Kerkhoff, C., Künsch, H.R. and Schär, C., 2014. Assessment of bias assumptions for climate models. *Journal of Climate*, *27*(17), pp.6799-6818.

King, L.M., McLeod, A.I. and Simonovic, S.P., 2015. Improved weather generator algorithm for multisite simulation of precipitation and temperature. *JAWRA Journal of the American Water Resources Association*, *51*(5), pp.1305-1320.

Krakauer, N.Y., 2014. Stakeholder-Driven Research for Climate Adaptation in New York City. In *New Trends in Earth-Science Outreach and Engagement* (pp. 195-207). Springer, Cham.

Krakauer, N.Y. and Fekete, B.M., 2014. Are climate model simulations useful for forecasting precipitation trends? Hindcast and synthetic-data experiments. *Environmental Research Letters*, *9*(2), p.024009.

Lall, U. and Sharma, A., 1996. A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, *32*(3), pp.679-693.

Lewis, S.C. and King, A.D., 2017. Evolution of mean, variance and extremes in 21st century temperatures. *Weather and climate extremes*, *15*, pp.1-10.

Lima, C.H., Lall, U., Troy, T.J. and Devineni, N., 2015. A climate informed model for nonstationary flood risk prediction: Application to Negro River at Manaus, Amazonia. *Journal of Hydrology*, *522*, pp.594-602.

Lins, H.F. and Cohn, T.A., 2011. Stationarity: wanted dead or alive? 1. *JAWRA Journal of the American Water Resources Association*, *47*(3), pp.475-480.

Mannshardt-Shamseldin, E.C., Smith, R.L., Sain, S.R., Mearns, L.O. and Cooley, D., 2010. Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data. *The Annals of Applied Statistics*, *4*(1), pp.484-502.

Milly, P.C., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P. and Stouffer, R.J., 2008. Stationarity is dead: Whither water management?, *Science*, *319*(5863), pp.573-574.

Namias, J., 1966. Nature and possible causes of the northeastern United States drought during 1962–65. *Mon. Wea. Rev*, *94*(9), pp.543-554.

National Research Council, 1999. *Improving American river flood frequency analyses* (chap. 4). National Academies Press.

Onof, C., Chandler, R.E., Kakou, A., Northrop, P., Wheater, H.S. and Isham, V., 2000. Rainfall modelling using Poisson-cluster processes: a review of developments. *Stochastic Environmental Research and Risk Assessment*, *14*(6), pp.384-411.

Oriani, F., Straubhaar, J., Renard, P. and Mariethoz, G., 2014. Simulation of rainfall time series from different climatic regions using the direct sampling technique. *Hydrology and Earth System Sciences*, *18*, pp.3015-3031.

Pierce, D.W., Cayan, D.R. and Thrasher, B.L., 2014. Statistical downscaling using localized constructed analogs (LOCA). *Journal of Hydrometeorology*, *15*(6), pp.2558-2585.

Razavi, S., Elshorbagy, A., Wheater, H. and Sauchyn, D., 2015. Toward understanding nonstationarity in climate and hydrology through tree ring proxy records. *Water Resources Research*, *51*(3), pp.1813-1830.

Seager, R., Pederson, N., Kushnir, Y., Nakamura, J. and Jurburg, S., 2012. The 1960s drought and the subsequent shift to a wetter climate in the Catskill Mountains region of the New York City watershed. *Journal of Climate*, *25*(19), pp.6721-6742.

Sharif, M. and Burn, D.H., 2007. Improved k-nearest neighbor weather generating model. *Journal of Hydrologic Engineering*, *12*(1), pp.42-51.

Sharma, A., Tarboton, D.G. and Lall, U., 1997. Streamflow simulation: A nonparametric approach. *Water resources research*, *33*(2), pp.291-308.

Shaw, E., 2014. Hydrology in practice. In (chap. 15). CRC Press.

Smith, K., Strong, C. and Rassoul-Agha, F., 2017. A new method for generating stochastic simulations of daily air temperature for use in weather generators. *Journal of Applied Meteorology and Climatology*, *56*(4), pp.953-963.

Stakhiv, E.Z., 2011. Pragmatic Approaches for Water Management Under Climate Change Uncertainty 1. *JAWRA Journal of the American Water Resources Association*, *47*(6), pp.1183-1196.

Steinschneider, S., Wi, S. and Brown, C., 2015. The integrated effects of climate and hydrologic uncertainty on future flood risk assessments. *Hydrological Processes*, *29*(12), pp.2823-2839.

Taner, M.Ü., Ray, P. and Brown, C., 2017. Robustness-based evaluation of hydropower infrastructure design under climate change. *Climate Risk Management*, *18*, pp.34-50.

Trenberth, K., Covey, C., Dai, A. and Fasullo, J., 2018. *Final Report on" Collaborative Research to Narrow Uncertainties in Precipitation and the Hydrological Cycle in Climate Models"* (No. DOE-UCAR-12602). National Center for Atmospheric Research, Boulder, CO (United States).

Trzaska, S. and Schnarr, E., 2014. A review of downscaling methods for climate change projections. *United States Agency for International Development by Tetra Tech ARD*, pp.1-42.

van Oldenborgh, G.J., Reyes, F.D., Drijfhout, S.S. and Hawkins, E., 2013. Reliability of regional climate model trends. *Environmental Research Letters*, *8*(1), p.014055.

Villarini, G., Serinaldi, F., Smith, J.A. and Krajewski, W.F., 2009. On the stationarity of annual flood peaks in the continental United States during the 20th century. *Water Resources Research*, *45*(8).

Wang X, Li J, Sun C, Liu T. NAO and its relationship with the Northern Hemisphere mean surface temperature in CMIP5 simulations. *Journal of Geophysical Research: Atmospheres*. 2017 Apr 27;122(8):4202-27.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O. and Schewe, J., 2014. The inter-sectoral impact model intercomparison project (ISI–MIP): project framework. *Proceedings of the National Academy of Sciences*, *111*(9), pp.3228-3232.

Westra, S., Alexander, L.V. and Zwiers, F.W., 2013. Global increasing trends in annual maximum daily precipitation. *Journal of climate*, *26*(11), pp.3904-3918.

Wilks, D.S. and Wilby, R.L., 1999. The weather generation game: a review of stochastic weather models. *Progress in physical geography*, *23*(3), pp.329-357.

Yates, D., Gangopadhyay, S., Rajagopalan, B. and Strzepek, K., 2003. A technique for generating regional climate scenarios using a nearest-neighbor algorithm. *Water Resources Research*, *39*(7).

Yiou, P., 2014. Anawege: a weather generator based on analogues of atmospheric circulation. *Geoscientific Model Development*, *7*(2), pp.531-543.

Young, K.C., 1994. A multivariate chain model for simulating climatic parameters from daily data. *Journal of Applied Meteorology*, *33*(6), pp.661-671.