atmosphere

MDPI

Type of the Paper (Article, Review, Communication, etc.)

Reanalysis Product-Based Nonstationary Frequency Analysis for Estimating Extreme Design Rainfall

Dong-IK Kim [1,*], Dawei Han [1] and Taesam Lee [2,*]

[1] Office 2.17, Queens Bulidng, Water and Environment Research Group, Department of Civil Engineering, University of Bristol, Bristol BS8 1TL, UK; d.han@bristol.ac.uk
[2] Department of Civil Engineering, ERI, Gyeongsang National University, Jinju 501, Korea
* Corresponding: dk15461@bristol.ac.uk (D.-I.K.); tae3lee@gnu.ac.kr (T.L.)

Supplementary: Description of Mathematical Background

1. Bias Correction

Although the century-long reanalysis precipitation data described above adopt observational data when modeling, the modeled data still include substantial biases. Bias corrections should be applied preliminarily to the model values before further hydrologic applications. In the current study, we first carried out bias corrections by a quantile mapping (QM) approach, typically adopted in bias correction studies (Cannon et al. 2015; Kim et al. 2015; Rabiei and Haberlandt 2015; Eum and Cannon 2017; Li et al. 2017). Conceptually, the QM method reduces errors by fitting a cumulative distribution of the modeled data to a cumulative distribution of the observations via a transfer function as follows (Teutschbein and Seibert 2012; Rabiei and Haberlandt 2015):

$$\hat{x}_m = F_o^{-1}[F_m(x_m)] \quad (1)$$

Here, $\hat{x}_m$ and x_m represent the bias-corrected values and the modeled data, respectively, and F_m and F_o indicate the cumulative distribution functions (CDFs) of the modeled and observed data, respectively. To avoid large amounts of noise corresponding to extreme values (Eum and Cannon 2017; Volosciuk et al. 2017; Maraun and Widmann 2018), we only consider parametric QM methods in the current study.

In the estimation of design rainfalls, a BM approach using AMRs is commonly adopted. To implement a reanalysis product-based BM approach, we directly improved the uncorrected AMRs by using QM without considering other rainfall data. This approach can reduce the error more efficiently than correcting the entire rainfall series (Li et al. 2017). To determine the best-fitting transfer function, we applied three representative distributions, namely, gamma, Gumbel, and GEV distributions, to correct the biases of the AMRs; these distributions are commonly employed in hydrologic applications and in the analysis of extremes (Koutsoyiannis 2004; Wilson and Toumi 2005; Kim et al. 2015; Rabiei and Haberlandt 2015). As a QM approach is based on a one-to-one relationship, we matched the 48 weather stations with the closest grid points of ERA-20c and 20CR, and bias correction parameters were collected at each station. Here, we assumed that the difference in the spatial resolution between the datasets can be ignored. One major issue in the bias correction of climate model data is how to correct the model values beyond the temporal range of the observations. The conventional approach of the QM algorithm was implemented under the assumption that the climate records were stationary for the whole projected period (Teutschbein and Seibert 2012; Cannon et al. 2015). More specifically, the CDFs of the observations and the modeled data are estimated using records from a reference period (i.e., a historical period or calibration period), and the model data for the whole projected period are applied to a correction factor as follows:

Citation: Kim, D.-I.; Han, D.; Lee, T. Reanalysis-Product-Based Nonstationary Frequency Analysis for Estimating Extreme Design Rainfall. *Atmosphere* **2021**, *12*, 191. https://doi.org/10.3390/atmos12020191

Academic Editor: Alexandre Ramos
Received: 14 January 2021
Accepted: 27 January 2021
Published: 31 January 2021

$$\hat{x}_{m,p} = F_{o,r}^{-1}[F_{m,r}\{x_{m,p}(t)\}] \quad (2)$$

Here, $F_{o,r}$ and $F_{m,\mathrm{r}}$ are the CDFs of the observed and modeled for the reference period (denoted by *r*), respectively, while $\hat{x}_{x,p}(t)$ and $x_{m,p}$ are the bias-corrected and uncorrected (or modeled) data, respectively, at time *t* during the whole projected period (denoted by *p*). In this concept, the years from 1974 to 2010 were set as the reference period, while the whole period from 1900 to 2010 was considered the projected period. For the stationary quantile mapping (SQM) scheme, there exist some extreme values beyond the range of the reference period that may overestimate the bias-correction results, so an appropriate extrapolation scheme should be considered for those values (Themeßl et al. 2012; Eum and Cannon 2017; Li et al. 2017). In the current study, we applied a constant extrapolation scheme, which applies the correction values at the lowest and highest quantiles of the calibration range, as suggested by Themeßl et al. (2012), to the events beyond the range of reference data; in constast, the AMRs within the range were corrected by parametric approaches based on three different distributions, the GEV, gamma, and Gumbel distributions. Note that the SQM approaches with the GEV, gamma, and Gumbel distributions are abbreviated hereafter as gevSQM and gamSQM and gumSQM, respectively.

One major problem in the SQM approach is that it ignores time-dependent characteristics, such as long-term trends. To address this issue, several approaches have been tested, such as the detrended quantile mapping and the quantile delta approach (Li et al. 2010; Bürger et al. 2013; Cannon et al. 2015; Miao et al. 2016; Eum and Cannon 2017). Among these approaches, we applied the quantile delta mapping (QDM) method suggested by Cannon et al. (2015) because this approach can preserve the changes not only in the mean but also in the extremes for the modeled data. In QDM, long-term trends in data are preserved by superimposing the relative changes in the quantiles between the reference period and the projected period, which are set to the same length. Thus, we first set the reference period to 1974-2010 (same as in SQM) and then divided the preceding projected period (1900-1973) into two periods, 1900-1936 and 1937-1973, to ensure that the lengths of these intervals were equal to the length of the reference period. Consequently, the reanalysis daily precipitations were divided into three time periods with the same length (1900-1936, 1937-1973, 1974-2010), and the raw models in each time period were improved by employing the QDM principle as follows (Cannon et al. 2015; Eum and Cannon 2017):

$$\tau_{m,p}(t) = F_{m,p}^{(t)}\left[x_{m,p}(t)\right], \quad \tau_{m,p}(t) \in \{0,1\} \quad (3)$$

$$\Delta_m(t) = \frac{F_{m,p}^{-1}\left(\tau_{m,p}(t)\right)}{F_{m,r}^{-1}\left(\tau_{m,p}(t)\right)} = \frac{x_{m,p}(t)}{F_{m,r}^{-1}[F_{m,p}(x_{m,p}(t))]} \quad (4)$$

$$\hat{x}_{m,p} = F_{o,r}^{-1}[\tau_{m,p}(t)] \times \Delta_m(t) = F_{o,r}^{-1}[F_{m,p}\{x_{m,p}(t)\}] \times \Delta_m(t) \quad (5)$$

Here, (1) $\tau_{m,p}(t)$ is the nonexceedance probability associated with the value at time *t* for a preceding period beyond the observation period (1900-1936 or 1937-1973); (2) $F_{m,r}$ and $F_{m,p}$ are the CDFs of the modeled for the reference period (1974-2010) and a projected period, respectively; and (3) $\Delta_m(t)$ represents the relative change in the quantiles between a preceding period and the reference period (1974-2010). Similar to the SQM approach, we applied three distributions, the GEV, gamma and Gumbel distributions, to estimate the CDFs in Eqs. 3 through 5, and the QDM schemes with these curves were named gevQDM, gamQDM, and gumQDM, respectively. A more specific description of QDM can be found in Cannon et al. (2015).

For the selection of the distribution, the root mean square error (RMSE) and Nash-Sutcliffe efficiency (NSE) have been commonly used for hydrological studies, as shown in Eqs. 6 and 7, respectively:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i^{\mathrm{obs}} - Y_i^{sim}\right)^2}{n}} \quad (6)$$

$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}\left(Y_i^{obs} - Y_i^{sim}\right)^2}{\sum_{i=1}^{n}\left(Y_i^{obs} - Y^{mean}\right)^2}\right] \quad (7)$$

Here, Y_i^{obs} and Y_i^{sim} represent the i-th observed and modeled values, respectively, among n data, while Y^{mean} denotes the mean of the observations. For the NSE, the dataset accuracy improves as the efficiency approaches 1, while a lower RMSE indicates better performance. In this analysis, we compared the bias-corrected AMRs obtained by using the three QM approaches with the observations from all 48 stations for the reference period (1974-2010). As the results of the QDM approaches were identical to those of the SQM schemes for the reference period, the QDM results were used to evaluate the performances of the three different curves for the bias correction scheme.

2. Detecting nonstationarity: Long-term trend test

As mentioned in the Introduction, the conventional bias correction approach is based on the condition of stationarity for climate model records, but the real climate may be nonstationary in terms of its century-long trends. To determine the significance of the AMR trend over South Korea, we evaluated the long-term trends of both the observations and the bias-corrected reanalysis data. For these trend tests, a nonparametric method, the Mann-Kendall test, was applied in this study. The significance of trends was evaluated by comparing the test statistic Z with the standard normal variate at the desired level of significance (Hamed and Rao 1998). When $|Z| > Z_{1-\alpha/2}$ for the standard normal deviate $Z_{1-\alpha/2}$ with the significance level α (= 0.05 in the current study), the null hypothesis is rejected, and a significant trend is detected in the time series. For the slope, the Theil-Sen approach (Theil 1950; Sen 1968), defined by the median among the ranked slope estimates, is applied as follows :

$$\beta = Med\left(\frac{x_j - x_i}{j - i}\right), \quad \forall i < j \quad (8)$$

where a positive value of β indicates an increasing trend over time and vice versa. The advantage of this method is that it is less sensitive to outliers or extreme values than the least-square method (Shadmani et al. 2012; Sayemuzzaman and Jha 2014).

We first analyzed the trends of the AMRs taken from both the observed data and the bias-corrected reanalysis data for the reference period (1974-2010). To estimate the nonstationarity over the 20th century, the century-long trends of the bias-corrected AMRs from 1900 to 2010 were also detected.

3. Rainfall Frequency Analysis under the condition of nonstationarity

In hydrological models, time-varying parameter schemes have been commonly adopted for nonstationarity analysis in hydrometeorological applications (Leclerc and Ouarda 2007; Ouarda and El-Adlouni 2011; Panagoulia et al. 2014; Salas and Obeysekera 2014; Du et al. 2015; Son et al. 2017). As the GEV family is typically applied for estimating intensity-duration-frequency relationships in practice, we applied a GEV distribution with the time-varying location parameter (μ_t), while the scale (σ) and shape (ξ) parameters

were set as constants. The location parameter is assumed to be the time-dependent linear function described in Eq. 9, and under the condition of nonstationarity, the CDF of the GEV distribution is described as Eq. 10.

$$\mu_t(t) = \mu_s t + \mu_i \tag{9}$$

$$F_z(z, \theta_t) = exp\left\{-\left[1 + \xi\left(\frac{z - \mu_t}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \tag{10}$$

Here, μ_i and μ_s are the intercept and slope of the location parameter, respectively, and $\theta_t = \{\mu_t, \sigma, \xi\}$ represents the time-varying parameter set of the GEV distribution.

To quantify the parameters for the GEV curve under the condition of nonstationarity, we applied the Bayesian principle suggested by Cheng et al. (2014). In this scheme, numerous parameter sets are estimated from the joint posterior distribution using the differential evolution Markov chain (DE-MC), which is based on the genetic algorithm differential evolution for global optimization with the Markov chain Monte Carlo (MCMC) principle (Cheng et al. 2014). More specifically, the posterior distribution, $p(\boldsymbol{\theta}|\mathbf{R})$, of the parameter vector ($\boldsymbol{\theta}$) is described as follows:

$$p(\boldsymbol{\theta}|\mathbf{R}) = \frac{p(\boldsymbol{\theta}, \mathbf{R})}{p(\mathbf{R})} = \frac{p(\mathbf{R}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{R})} = \frac{p(\mathbf{R}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\mathbf{R}|\boldsymbol{\theta})d\theta} \propto p(\mathbf{R}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{11}$$

where $\mathbf{R}$ is the vector of the bias-corrected AMRs, $p(\mathbf{R}|\boldsymbol{\theta})$ is the likelihood function, and $p(\mathbf{R})$ and $p(\boldsymbol{\theta})$ are the marginal distribution and prior distribution, respectively. In this approach, normal distributions are used for the priors of parameters, and the prior distributions for all parameters are assumed to be independent (Cheng et al. 2014). The joint posterior distribution function $p(\boldsymbol{\theta}|\mathbf{R})$ can be formulated by combining the GEV likelihood function and prior distribution as follows:

$$p(\boldsymbol{\theta}|\mathbf{R}) \propto \prod_{i=1}^{n} GEV(\mathbf{R}|\mu_t, \sigma, \xi) \tag{12}$$

The posterior distribution for the parameter vector was obtained by maximizing the joint posterior distribution illustrated in Eq. 12 via an MCMC scheme. Further information can be found in Cheng et al. (2014).

With these time-varying parameter chains, the next step is to estimate the return period for a given design quantile under the condition of nonstationarity. Numerous studies have addressed the nonstationary assumption, for which two main approaches were developed: (1) the expected waiting time (EWT) method and (2) the expected number of exceedance (ENE) method (Obeysekera and Salas 2014, 2016; Salas and Obeysekera 2014; Du et al. 2015; Read and Vogel 2015; Salas et al. 2018).

The EWT interpretation starts from estimating the probability for the first occurrence exceeding a design quantile (z_{q0}). Under the condition of nonstationarity, the exceedance probabilities (p_t) vary with time, and the first occurrence exceeding the design quantile (z_{q0}) at time x is described as follows (Salas and Obeysekera 2014; Du et al. 2015; Salas et al. 2018):

$$\begin{aligned} f(x) = P(X = x) &= (1 - p_1)(1 - p_2)(1 - p_3)\cdots(1 - p_{x-1})p_x \\ &= p_x \prod_{t=1}^{x-1}(1 - p_t), \quad x = 1, 2, \cdots, \infty \end{aligned} \tag{13}$$

The expected waiting time for the first event exceeding z_{q0}, i.e., the return period (T), is obtained as follows:

$$T = E(X) = \sum_{x=1}^{\infty} xf(x) = \sum_{x=1}^{\infty} xp_x \prod_{t=1}^{x-1}(1 - p_t) \quad (14)$$

Unlike the EWT scheme, the ENE approach focuses on the expected number of exceedances over the design life T. The number of events (Y) exceeding the design rainfall (z_{q0}) in T years can be expressed as follows (Du et al. 2015; Salas et al. 2018):

$$Y = \sum_{j=1}^{T} p_t \quad (15)$$

The return period (T) for the first event exceeding the design quantile (z_{q0}) can be numerically estimated by applying $Y = 1$ in Eq. 15. Here, the exceedance probability (p_t) corresponding to the design quantile (z_{q0}) is expressed for the GEV distribution as follows:

$$p_t = 1 - exp\left\{-\left[1 + \xi\left(\frac{z_{q0} - \mu_t}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \quad (16)$$

Both EWT and ENE are applicable for nonstationary events. However, the EWT approach has the drawback of requiring infinitely many (or as many as possible) future exceedance probabilities to numerically solve the problem (Du et al. 2015; Salas et al. 2018). For this reason, we applied the ENE scheme to estimate the design quantile (z_{q0}) with a return period ranging from 10 years to 200 years.

In the proposed approach, four parameters (μ_s, μ_i, σ and ξ) of the GEV distribution are required to estimate the intensity-duration-frequency relationships under the condition of nonstationarity. By using the Bayesian principle in Eqs. 11 and 12, we collected the time-varying parameter sets ($\mu_{s,m}$, $\mu_{i,m}$, σ_m and ξ_m) based on the bias-corrected AMRs of ERA-20c and 20CR from 1900 to 2010. However, as these bias-corrected values may still have errors of a certain magnitude, estimating the future risk by these model parameters may also misrepresent that future risk. On the other hand, the conventional approach based on observations may not sufficiently represent the long-term changes in the AMRs due to the lack of data in certain regions including South Korea. For this reason, we estimated the nonstationary design rainfall by using all four parameter sets ($\mu_{s,m}$, $\mu_{i,m}$, σ_m and ξ_m) derived from the bias-corrected AMRs of ERA-20c and 20CR at 48 stations. More specifically, we explored the design quantiles with return periods ranging from 10 years to 200 years by applying the median and 90% confidence interval (CI) of the four parameter chains generated by Eqs. 11 and 12. Note that in this analysis, we considered 2011 the beginning year (t_1) for the target return period in Eq. 15. A flow chart for the nonstationary analysis is illustrated in Figure S1.

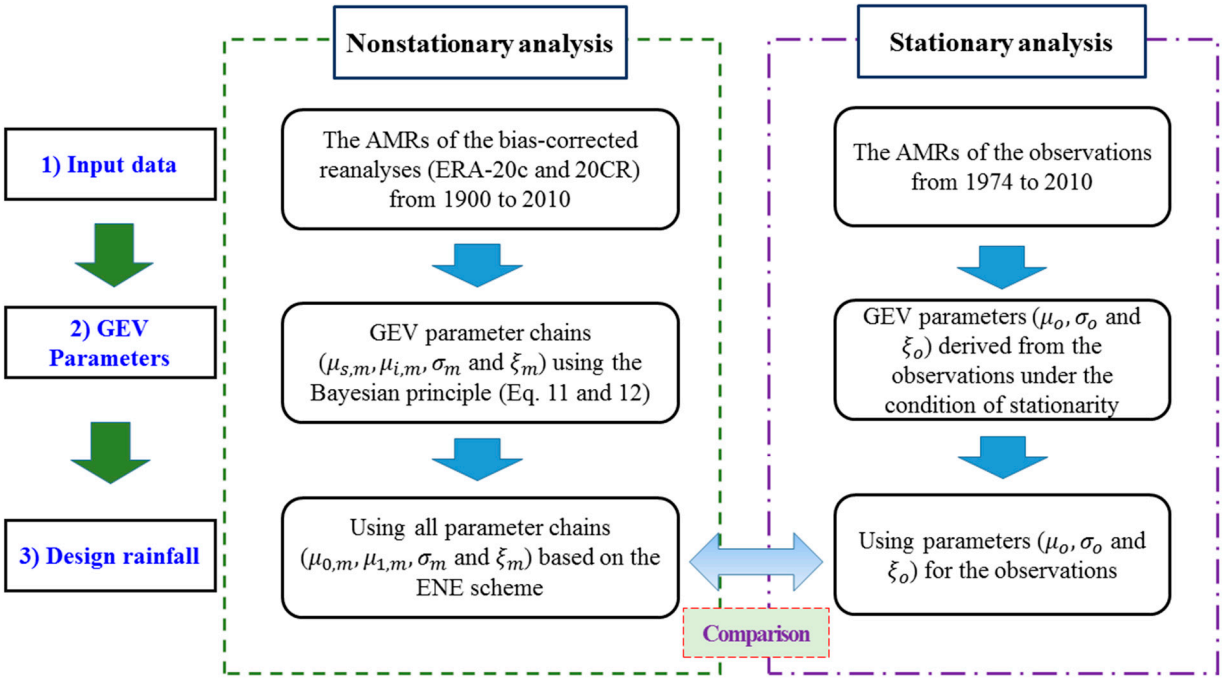

Figure S1. A flow chart for estimating design rainfall under the conditions of nonstationarity and stationarity.

Considering the condition of nonstationarity at all 48 stations, we finally explore the spatial change in design rainfall with a 100-year return period over South Korea. After creating contour maps of design rainfalls for both the nonstationary model and the stationary model based on a scattered data interpolation method in MATLAB (Amidror 2002), we spatially assessed the relative difference (RD, %) between the modeled design rainfalls and the conventional values as follows:

$$RD(\%) = \frac{D_r^{mod} - D_p^{obs}}{D_r^{obs}} \times 100 \quad (17)$$

Here, D_r^{obs} represents the design rainfall using the observed AMRs for the reference period (i.e., 1974-2010), while D_p^{mod} indicates the design rainfalls based on the nonstationary interpretation. Here, the design rainfalls for the nonstationary model were estimated from the median values of the parameter chains for the GEV distribution.