Streamflow estimation at partially gaged sites using multiple dependence conditions via vine copulas

Kuk-Hyun Ahn[1]

October 2020

[1]Assistant Professor, Department of Civil and Environmental Engineering, Kongju National University, Cheon-an, South Korea; *Corresponding author;* e-mail: ahnkukhyun@gmail.com

ABSTRACT

Reliable estimates of missing streamflow values are relevant for water resources planning and management. This study proposes a multiple dependence condition model via vine copulas for the purpose of estimating streamflow at partially gaged sites. The proposed model is attractive in modeling the high dimensional joint distribution by building a hierarchy of conditional bivariate copulas when provided a complex streamflow gage network. The usefulness of the proposed model is firstly highlighted using a synthetic streamflow scenario. In this analysis, the bivariate copula model and a variant of the vine copulas are also employed to show the ability of the multiple dependence structure adopted in the proposed model. Furthermore, the evaluations are extended to a case study of 54 gages located within the Yadkin-Pee Dee River Basin, the eastern U. S. Both results inform that the proposed model is better suited for infilling missing values. To be specific, the proposed multiple dependence model shows the improvement of 9.2 % on average compared to the bivariate model from the historical case study. The performance of the vine copula is further compared with six other infilling approaches to confirm its applicability. Results demonstrate that the proposed model produces more reliable streamflow estimates than the other approaches. In particular, when applied to partially gaged sites with sufficient available data, the proposed model clearly outperforms the other models. Even though the model is illustrated by a specific case, it can be extended to other regions with diverse hydro-climatological variables for the objective of infilling.

Keywords: **vine copulas, multiple dependence condition model, infilling approach, and streamflow estimation at partially gaged site**

1. Introduction

Hydrological observation records covering long-term periods are instrumental in water resources planning and management including the design of flood defense systems and irrigation water management (Aissia et al., 2017; Beguería et al., 2019). However, available streamflow data is often limited due to several situations like equipment failures, budgetary cuts, and natural hazards (Kalteh and Hjorth, 2009). Missing data is particularly observed in remote catchments where equipment failures are repaired only after significant delays following extreme events, which can be crucial for hydrological frequency analysis. Hence, hydrologists often rely on simulated sequences to infill missing data in partially gaged catchments (Booker and Snelder, 2012) by using two primary modeling approaches such as: (1) process-based models (i.e., estimating streamflow based on a conceptual understanding of hydrological processes), and (2) transfer-based statistical models (i.e., transferring information from gaged to ungagged catchments) (Farmer and Vogel, 2016). This paper focuses on the latter, which estimates historical daily streamflow at inadequately and partially gaged sites by the means of a statistical relationship.

Over the past few decades, a variety of statistical models including simple drainage area scaling (Croley and Hartmann, 1986), spatial interpolation technique (Pugliese et al., 2014), regression model (Beauchamp et al., 1989) and flow duration curves (FDCs; Hughes and Smakhtin, 1996), have been developed. In particular, the flow duration curve method has been regarded as one of the most trustworthy regionalization approaches (Archfield and Vogel, 2010; Boscarello et al., 2016; Castellarin et al., 2004; Li et al., 2010; Mendicino and Senatore, 2013). If the target watershed is completely ungaged, FDCs can be established using regression models to regionalize the parameter sets of defined distributions (e.g., Ahn and Palmer, 2016a; Blum et al., 2017) or to regionalize a set of primary quantiles (Cunderlik and Ouarda, 2006; Schnier and Cai, 2014; Zaman et al., 2012). On the other hand, if the target watershed is poorly or partially gaged, FDC models are built using the following four steps: (1) estimating non-exceedance probability for recorded streamflow from the target watershed of interest; (2) selecting one or multiple donor watersheds for the target watershed; (3) transferring the time-series of non-exceedance probability from the donor watersheds for missing streamflow values; and (4) converting corresponding streamflow values back from the transferred non-exceedance probability. When FDCs are utilized for partially gaged watersheds, how the donor watersheds are selected (step 2) and how the probabilities are transferred from the donor watersheds (step 3) are fairly crucial in the FDC framework.

Many studies have developed diverse approaches for steps 2 and 3 in FDC modelling. While the basic formulation is that non-exceedance probabilities of the target site are transferred by those at the single donor site, a weighted average of non-exceedance probability from the selected donor sites has been suggested by Smakhtin (1999) instead. In addition, Farmer (2015) adopted a kriging model to regionalized daily standard (i.e., z-scored) probabilities based on non-exceedance probabilities from many donors in a region, using the quantile function of a standard normal distribution. Although these studies are promising, the joint distribution of non-exceedance probability between the target and donor watersheds is modeled based on a Gaussian assumption which cannot properly permit different percentile values such as extremes that have different spatial dependence structures from donor sites. To circumvent this limitation, Worland et al. (2019) suggested the copula theory after showing that a unifying framework of copulas is equivalent to that of FDC (i.e., estimations of the conditional probabilities at the target watershed given known values at the donors).

Increasing attention has been received to copulas in the field of hydrology, with applications in flood frequency analysis, drought risk analysis, and multi-site streamflow generations (Ahn and Palmer, 2016b; Ariff et al., 2012; Chen et al., 2015; Daneshkhah et al., 2016; Fu and Butler, 2014). Copulas are effective mathematical functions that are capable of combining univariate marginal distribution functions of random variables into their joint cumulative distribution function and allow representation of diverse dependence structures between these random variables corresponding to their family members (Sklar, A., 1959). For example, Fu and Butler (2014) showed that the Gumbel copula performs well in representing multiple flooding characteristics as compared to the other copulas from the Archimedean family, namely the Clayton and Frank copulas. To estimate streamflow (i.e., infilling missing data) at poorly and partially gaged sites, Worland et al. (2019) have developed bivariate copulas with an Archimedean copula, but limited their application to a single donor. Albeit the limitation, their bivariate copulas may be acceptable since the higher dimension of copulas is not rich enough to model all possible mutual dependencies among multisite donors (see Karmakar and Simonovic, 2009 for details). Hao and Singh (2013) also describe that multivariate copulas are incapable of modeling multisite data exhibiting complex patterns of dependence.

However, if the theoretical limitation of a multivariate copula is mitigated, dependency information from multiple donor sites may allow more reliable predictions of regionalized streamflow. Vine copulas, also known as pair copulas, offer a far efficient way to construct higher dimensional dependence (Bedford et al., 2002; Joe, 2014). They have hierarchical structures that sequentially apply bivariate copulas as the building local blocks for constructing a higher dimensional copula. The high flexibility of vine copulas enables modeling a wide range of complex data dependencies. In particular, Aas et al. (2009) have popularized two classes of vine copulas, canonical vines (C-vines) and drawable vines (D-vines) by allowing diverse pair-copula families such as the bivariate Student-t copula and bivariate Clayton copula. After the seminal paper, those two vines have been used in many fields including economics (Arreola Hernandez et al., 2017; Zimmer, 2015), finance (Dissmann et al., 2013; Lu, 2013), and engineering (Bhatti and Do, 2019; Erhardt et al., 2015; Xu et al., 2017). Similarly, a few studies have used vine copulas in hydrologic applications with diverse purposes (Daneshkhah et al., 2016; Liu et al., 2015; Vernieuwe et al., 2015; Shafaei et al., 2017) although they have not been introduced to infill missing data.

Based on the usefulness of vine copulas, Kraus and Czado (2017) have developed a promising algorithm that sequentially fits such a D-vine copula model ($\mathcal{M}_{\mathrm{Kraus}}$). The algorithm adds covariates to the model with the objective of maximizing a conditional likelihood and stops adding covariates to the model when none of the remaining covariates can significantly increase the model's conditional likelihood. While it is promising, one challenge that can arise but has not been previously discussed is overfitting when covariates are correlated with each other. In this situation, the model may adopt ineffective covariates and eventually leads to poor predictions. In particular, for the purpose of infilling, streamflow values at the target site are often correlated by those of many donors. Although the structure of $\mathcal{M}_{\mathrm{Kraus}}$ is potentially favorable to estimate streamflow, modified model procedure is required to determine the most influential covariates.

This study forwards two novel contributions to infill missing data in the field of hydrology: (1) a D-vine copula-based model is introduced to estimate streamflow for poorly and partially gaged watersheds and (2) the existing model ($\mathcal{M}_{\text{Kraus}}$) is further improved by incorporating a new procedure to determine the optimal number of donor sites (namely $\mathcal{M}_{\text{Dvine}}$). First, synthetic data are generated to compare $\mathcal{M}_{\text{Kraus}}$ and $\mathcal{M}_{\text{Dvine}}$. In this analysis, bivariate copulas (namely $\mathcal{M}_{\text{Bicop}}$) is also employed to demonstrate the usefulness of a high dimensional joint dependence structure. Afterwards, a real infilling example is utilized to compare the proposed vine-based model with six other streamflow-transfer models adopted in literatures.

2. Methodology

2.1 D-vine copulas

A copula C is k-variate cumulative distribution function on $[0,1]^k$ with all uniform margins. The C can be understood as a function that links the marginal cumulative distributions $(F_1,\dots,F_k)$ to form a joint distribution F. The C associated with joint distribution F is a distribution function $C\colon[0,1]^k \rightarrow [0,1]$ such that, for all streamflow vector $\boldsymbol{q} = (q_1,\dots,q_k)^T$, the C satisfies:

$$F(q_1,\dots,q_k) = C(F_1(q_1),\dots,F_k(q_k)) \qquad \text{Eq. (1)}$$

where C is unique if $F_1, \dots, F_k$ are continuous.

Based on Sklar's theorem (Sklar, A., 1959), a multivariate distribution function is a composition of a set of marginal distributions; thus, equation (1) can be expressed in terms of densities,

$$f(q_1, \dots, q_k) = [\prod_{i=1}^{k} f_i(q_i)] c(F_1(q_1), \dots, F_k(q_k)) \qquad \text{Eq. (2)}$$

where c is a k-dimensional copula density acquired by partial differentiation of the copula C (i.e., $c\big(F_1(q_1), \dots, F_k(q_k)\big) := \frac{\partial^k}{\partial_1 \cdots \partial_k} C\big(F_1(q_1), \dots, F_k(q_k)\big)$) and $f_i(\cdot)$ is the marginal density corresponding to $F_i(\cdot)$.

Following Bedford and Cooke (2001), any copula density $c(F_1(q_1), \dots, F_k(q_k))$ can be decomposed into a product of $k(k-1)/2$ pair copula densities. Aas et al. (2009) adopted this idea and introduced the copula class of pair copula constructions (PCCs) known as vine copulas. These copulas are suitable to model various dependency structures. Vine structures established by $k(k-1)/2$ pair copulas are arranged in $k-1$ trees (Brechmann et al., 2013) and can be categorized as C-vines and D-vines (Liu et al., 2015). This study focuses on D-vines since they are more widely used in practice (Daneshkhah et al., 2016).

A D-vine is characterized by the ordering of its variables (see Figure 1). In the first tree, the dependence of the first and second variables, of the second and third, of the third and fourth, and so on, is modeled using pair-copulas. In the second tree, conditional dependence of the first and third given the second variable (i.e., $c_{1,3|2}(F(q_1|q_2), F(q_3|q_2))$), the second and fourth given the third (i.e., $c_{2,4|3}(F(q_2|q_3), F(q_4|q_3))$), and so on, is modeled. Similarly, pairwise dependencies of two variables are modeled in subsequent trees conditioned on those variables which lie between the two variables in the first tree (e.g., $c_{1,5|2,3,4}(F(q_1|q_2,q_3,q_4), F(q_5|q_2,q_3,q_4))$). The density of the k-dimensional D-vine can be computed as follows (Aas et al., 2009):

$$f(q_1,\dots,q_k) = \left[\prod_{i=1}^{k} f_i(q_i)\right] \times \prod_{j=1}^{k-1}\prod_{i=1}^{k-j} c_{j,j+i|(i+1):(j+i-1)}\left(F\left(q_i\middle|q_{i+1},\dots,q_{i+j-1,}\right), F\left(q_{i+j}\middle|q_{i+1},\dots,q_{i+j-1,}\right)\right) \quad \text{Eq. (3)}$$

where $c_{j,j+i|(i+1):(j+i-1)}$ indicates the bivariate copula densities.

For the five-dimensional D-vine copula as an example in Figure 1, the corresponding vine distribution has the joint density as follows:

$$f(q_1,\dots,q_5) = \left[\prod_{i=1}^{5} f_i(q_i)\right] c_{12}\cdot c_{23}\cdot c_{34}\cdot c_{45}\cdot c_{13|2}\cdot c_{24|3}\cdot c_{24|3}\cdot c_{35|4}\cdot c_{14|23}\cdot c_{25|34}\cdot c_{15|234} \quad \text{Eq. (4)}$$

where $c_{1,2}(F_1(q_1), F_2(q_2))$ is simply denoted as $c_{1,2}$.

As presented in equation (4), the conditional distribution functions and conditional bivariate copulas are required in vine copula modeling. The conditional distribution functions $F(q_{\jmath}|q_{\jmath+1}, \dots, q_{\jmath+j-1})$, also known as h-functions, in equation (4) can be addressed using the pair-copulas from lower trees by using equation (5). Let q_i be a conditional value of $q_{\jmath+1}, \dots, q_{\jmath+j-1}$ and $\boldsymbol{v} = \{q_{\jmath+1}, \dots, q_{\jmath+j-1}\} \backslash q_i$ the streamflow vector without q_i used in the following recursive relationship (Aas et al., 2009):

$$h(q_{\jmath}|\boldsymbol{v}) := F(q_{\jmath}|\boldsymbol{v}) = \frac{\partial C_{\jmath i|\boldsymbol{v}}(F(q_{\jmath}|\boldsymbol{v}), F(q_i|\boldsymbol{v}))}{\partial F(q_i|\boldsymbol{v})} \quad \text{Eq. (5)}$$

where the h-function is associated with the pair-copula $C_{\jmath i|\boldsymbol{v}}$.

More details about D-vines can be found in Bedford et al., (2002) and Czado (2010, 2019).

2.2 Algorithm of D-vine copula-based estimation ($\mathcal{M}_{\text{Dvine}}$)

Following Kraus and Czado (2017), a two-step estimation procedure is utilized for the prediction of the streamflow value at the target watershed. The algorithm ($\mathcal{M}_{\text{Dvine}}$) is developed using two library packages in the R programming language (Bevacqua, 2017; Schepsmeier et al., 2015).

Let q_k be the quantile of streamflow at the target watershed given the streamflow values $q_1, \ldots, q_{k-1}$ from the donor sites. In the first step, the marginal cumulative probabilities $F_k(q_k)$ and $F_j(q_j)$, $j = 1, \ldots, k-1$, are estimated using the semiparametric approach. To be specific, this study uses the continuous kernel smoothing estimator (Geenens, 2014), which is, given observed streamflow q_i^ζ, $\zeta = 1, \ldots, \xi$, at ith site, defined as $\widehat{F_i}(q_i) = \frac{1}{nh}\sum_{\zeta=1}^{\xi} \Omega(\frac{q_i - q_i^\zeta}{h})$. Here, $\Omega(q_i)$ is the "kernel" function with $\omega(\cdot)$ being a symmetric probability density function and h is the parameter controlling the smoothness of the final estimate. In this study, a Gaussian kernel is used for all $\omega(\cdot)$. The estimated cumulative probabilities are then employed to model the D-vine copula in the second step.

Next, to easily estimate conditional streamflow values at the target site, the D-vine copula is fitted with fixed order $F_k(q_k) - F_{\mathrm{I}_1}(q_{\mathrm{I}_1}) - F_{\mathrm{I}_2}(q_{\mathrm{I}_2}) - \ldots - F_{\mathrm{I}_{k-1}}(q_{\mathrm{I}_{k-1}})$, such that $F_k(q_k)$ is the first node in the first tree and the other orders of donors (I_1, …, I_{k-1}) are decided based on their correlations to the target site (i.e., $F_{\mathrm{I}_1}(q_{\mathrm{I}_1})$ showing the greatest correlations to $F_k(q_k)$). To build the D-vine copula model, five bivariate copulas (Gaussian, Student-t, Frank, Gumbel, and Clayton copulas) are considered as potential pair copulas (building blocks) although more families of Copulas such as extreme value copulas (EVC) are desirable. The five candidates may be sufficient to represent diverse dependence structures. For example, a Gaussian copula is proper when the non-exceedance probabilities between two watersheds are associated in the body of their distribution but are not asymptotically dependent in the both tails. On the other hand, a Gumbel copula may be appropriate for the situation wherein the non-exceedance probabilities exhibit tail dependence, where high flows are connected by same rainfall events but low flows are not related (e.g., due to regulation) (Salvadori and De Michele, 2004). Details of the five bivariate copulas are presented in the Supporting Information. Parameters for the five bivariate copulas are estimated based on Kendall rank-based correlation (ρ^{τ}) between sites. The optimal bivariate copula for each pair copula is determined based on the penalized likelihood function (i.e., AIC).

The final number (χ_k) of donor sites is further optimized under a cross-validation approach. In this approach, 80 % of the regional data are employed for model fitting; the other 20 %, for testing. Again, this procedure is conducted 5 times, each time using a different set of data for testing. As a measure for the model's fit, the root mean squared error (RMSE; equation (6)) from observed streamflow at the target site is utilized.

$$RMSE_{\chi k} = \sqrt{\frac{1}{\xi}\sum_{\zeta=1}^{\xi}(q_k - \hat{q}_k^{\chi})^2} \qquad \text{Eq. (6)}$$

Finally, conditional streamflow values at the target site can be estimated using the inverse form of the conditional distribution function (i.e., Eq. 5). To depict the ideas, a trivariate case (i.e., $\chi = 2$) is considered here. Based on the streamflow values at the donor sites (q_2, q_3), $\widehat{q_1}$ can be obtained using the conditional distribution function $h(q_1|q_2, q_3)$. For some fixed probabilities ϕ (e.g., $\phi = 0.1, \dots, 0.9$), $F_1(\widehat{q_1})$ is derived from $C_{1|2,3}$ using an explicit function:

$$C_{1|2,3}^{-1}\left(\phi|F_2(q_2),F_3(q_3)\right) = h_{1|2}^{-1}(h_{1|32}^{-1}(\phi|h_{2|1}(F_2(q_2)|F_1(q_1)))|F_1(q_1)) \qquad \text{Eq. (7)}$$

where $C_{1|2,3}^{-1}$ is the inverse of the copula function given the ϕ quantile curve of the copula (Liu et al., 2015; Xu and Childs, 2013). Therefore, the ϕth copula-based conditional quantile function of streamflow at the target site can be calculated as follows:

$$q_1(\phi|q_2q_3) = F_1^{-1}(C_{1|2,3}^{-1}\left(\phi|F_2(q_2),F_3(q_3)\right)) =$$
$$F_1^{-1}(h_{1|2}^{-1}(h_{1|32}^{-1}(\phi|h_{2|1}(F_2(q_2)|F_1(q_1)))|F_1(q_1))) \qquad \text{Eq. (8)}$$

Similarly, for the k-dimensional case, the ϕth copula-based conditional quantile function can be calculated along with streamflow at the k-1 donor sites. To acquire an estimate at the target site, 1000 samples from uniform distribution over the interval [0, 1] are generated using Monte Carlo simulations. In this study, the mean value of these generations is regarded as the best estimate.

3. Application

This study first explores the performance of $\mathcal{M}_{\mathrm{Dvine}}$ under synthetic example. In this analysis, $\mathcal{M}_{\mathrm{Bicop}}$ and $\mathcal{M}_{\mathrm{Kraus}}$ are also employed to show the usefulness of $\mathcal{M}_{\mathrm{Dvine}}$. For $\mathcal{M}_{\mathrm{Bicop}}$, the optimal bivariate copula is selected based on the AIC while the five bivariate copulas (Gaussian, Student-t, Frank, Gumbel, and Clayton copulas) are considered as its potential candidates. A brief description of two additional models are presented in the supporting information. After that, those three models are used for a real application to 54 stream gages located in a region of the eastern United States by estimating streamflow in partially gaged locations. Finally, seven infilling approaches (Table 1) are also utilized and evaluated in a cross-validated framework to evaluate the performance of the proposed model.

3.1 Synthetic simulation

Synthetic streamflow data are generated using controlled Monte Carlo experiment to explore how well the three copula-based models ($\mathcal{M}_{\text{Bicop}}$, $\mathcal{M}_{\text{Kraus}}$, $\mathcal{M}_{\text{Dvine}}$) provide streamflow predictions at the target site given a complex streamflow data in a pseudo gage network. In this analysis, a six-dimensional streamflow set (q_1^ζ, q_2^ζ, q_3^ζ, q_4^ζ, q_5^ζ, q_6^ζ), $\zeta = 1, \ldots, \xi = 2190$ (i.e. $\frac{2190}{365}$ = 6 years), is modelled using four bivariate copulas (Gaussian, Student-t, Flank, and Clayton copulas) and lognormal distributions for margins (see Figure 2).

The performance of each model is evaluated in a calibration-validation framework. First, synthetic streamflow data are generated for six-dimensional gage network. Then, φ years of data are randomly selected to be assumed known at the target gage, and the streamflow for the remaining 6-φ years of data is then estimated as missing values ($\varphi = 4$ in this analysis). This process is repeated 20 times to build an ensemble prediction. In particular, this study assumes the fifth streamflow data (i.e., q_5) to be predicted. In this assessment, two characteristics are considered to compare the three models: model prediction reliability and uncertainty quantification skill. Model prediction reliability is tested using the root mean squared error

(RMSE; Eq. 6) and Nash-Sutcliffe efficiency (NSE), which are further described in Section 3.4. Uncertainty quantification skill is judged by the ability of each model to build prediction intervals (PIs) that correctly bound predictions (see Section 3.4). Here, coverage probabilities, defined as the proportion of the time that true values occur into these PIs, are employed to show the usefulness of the proposed model.

3.2 Application to the Yadkin-Pee Dee River

The Yadkin-Pee Dee River Basin (Figure 3), covering around 18,700 km^2 and one of the largest river basins in North Carolina and South Carolina (Fisk, 2010), is used as real data to evaluate infilling ability. The basin flows from the northwestern corner of North Carolina near Blowing Rock and extends south by southeast, crossing the south-central border of North Carolina into South Carolina, with slightly more than half of its watershed in North Carolina. Most of the land covered within the basin is forested or used for agriculture although urban areas of the basin are expanding.

Daily streamflow data at 54 gages are gathered throughout the study region from web interface of the U.S. Geological Survey (USGS) National Water Information System (NWIS) (U.S. Geological Survey, 2018). The 54 gages are selected based on the following criteria: (1) all gages are recorded continuously for 15 years of daily streamflow over the period from January 2004 to December 2018, and (2) gages have non-zero daily values for the period in the first criterion since gages with streamflow values equal to zero require a more flexible modeling structure. Thus, it is common to model zero flows separately in regionalization studies. Based on the second criterion, this study discards 10 gage stations (not shown).

3.3 Intermodel comparison framework

A set of seven infilling approaches is used in the final assessment (see Table 1): (1) $\mathcal{M}_{\text{FDC−IDW}}$, (2) $\mathcal{M}_{\text{IDW−streamflow}}$, (3) $\mathcal{M}_{\text{Rho−streamflow}}$, (4) $\mathcal{M}_{\text{FDC−highestrho}}$, (5) $\mathcal{M}_{\text{DAR−streamflow}}$, (6) $\mathcal{M}_{\text{Kriging−streamflow}}$, and (7) $\mathcal{M}_{\text{Dvine}}$. This set of seven models is tested in a cross-validation framework under two different cases. The two cases consider situations wherein φ have values of 2 and 8 to represent relatively deficit- and sufficient-records for the target site. Similar to the comparative assessment to show the usefulness of the proposed copula-based model (see Section 3.1), each case is repeated 20 times by randomly selecting φ years over the applied period. The reliability of each model is evaluated using RMSE and NSE metrics over the validated four-year period randomly selected in the remaining data (i.e., 4 years in 15-φ years).

3.4 Error metrics and error decomposition

As presented in Sections 3.1 and 3.3, the root mean squared error (RMSE; Eq. 6) and Nash-Sutcliffe efficiency (NSE) are employed to evaluate prediction skills:

$$NSE = 1 - \frac{\sum_{\zeta=1}^{\xi}(\widehat{q^{\zeta}} - q^{\zeta})^2}{\sum_{\zeta=1}^{\xi}(q^{\zeta} - \overline{q^{\zeta}})^2} \qquad \text{Eq. (9)}$$

The NSE (RMSE) can range from −∞ to 1 (0 to ∞), with higher NSE (lower RMSE) implying better performance. Both metrics have been commonly used in hydrology analysis (Boyle et al., 2000).

Following derivations suggested in Gupta et al. (2009), the RMSE can be further decomposed into three components:

$$RMSE^2 = MSE = (\hat{\mu} - \mu)^2 + (\hat{\sigma} - \sigma)^2 + 2\sigma\hat{\sigma}(1 - r) \qquad \text{Eq. (10)}$$

where μ ($\hat{\mu}$) and σ ($\hat{\sigma}$) represent the average and standard deviation for the observed (estimated) streamflow, respectively, and r indicates the estimated correlation coefficient. The first component $(\hat{\mu} - \mu)^2$ is a measure of how well the average of the observed streamflow represents the average of the estimated streamflow; the second component $(\hat{\sigma} - \sigma)^2$ is a measure of how well the variance of the prediction represents the variance of the observed streamflow; and the third component $2\sigma\hat{\sigma}(1 - r)$ is dominated by the correlation and is defined as the "timing" component (Worland et al., 2019). Using these three defined components, their absolute contributions are explored in this study.

In addition, the accuracy of the uncertainty quantification skill is also evaluated for the copula-based models ($\mathcal{M}_{\text{Bicop}}$, $\mathcal{M}_{\text{Kraus}}$, $\mathcal{M}_{\text{Dvine}}$). To be specific, this study utilizes the PI coverage probability (PICP), which a common metric for this purpose (He et al., 2017; Niemierko et al., 2019). It provides the relative number of data points that fall between the defined bounds as expressed follows:

$$\text{PICP} = \frac{1}{\xi}\sum_{\zeta=1}^{\xi}\Theta_\zeta \ \text{ with } \ \Theta_\zeta = \begin{cases} 1, & if \quad q^\zeta \in [L^\zeta, U^\zeta] \\ 0, & \text{else} \end{cases} \qquad \text{Eq. (11)}$$

where Θ_ζ is the indicator variable if q^ζ is covered by the ζth PI defined by the lower bound L^ζ and upper bound U^ζ. This study examines the prediction accuracy of single quantiles. Therefore, the lower bound is defined as $L^\zeta = -\infty$ and $U^\zeta = \widehat{q^{\zeta,\varpi}}$ where ϖ is the estimated quantile at time ζ. Accordingly, the upper bound is not a constant, but is re-assigned. By subtracting the nominal confidence ϖ from PICP, the average coverage error (ACE) is obtained as follows:

$$\text{ACE} = \text{PICP} - \varpi \qquad \text{Eq. (12)}$$

The metric clearly indicates if the predicted quantile is underestimated (ACE < 0) or overestimated (ACE > 0) while taking small values around 0 for ideal case.

4. Results

4.1 Results for synthetic experiment

Prediction results from out-of-samples for the RMSE and NSE metrics are presented for the three copula-based models ($\mathcal{M}_{\text{Bicop}}$, $\mathcal{M}_{\text{Kraus}}$, $\mathcal{M}_{\text{Dvine}}$) in Table 2. The ACE scores are also described for $\varpi \in \{0.05, 0.10, 0.50, 0.90, 0.95\}$ in Table 3. When compared to the other models, $\mathcal{M}_{\text{Bicop}}$ achieves lower RMSE values in the right tail of the RMSE distribution over the validation periods, but severely underperforms the majority of the designed experiment, suggesting this model formulation relying on a single donor leads to poor predictions. $\mathcal{M}_{\text{Kraus}}$ provides higher RMSE values for all the RMSE distribution, particularly for the right tail of the RMSE distribution. The model utilizes streamflow data from all donors (i.e., five donor sites) although the first two gages (Gages 1 and 2) show insignificant associations to the target site ($r_{1,5} = 0.11$ and $r_{2,5} = 0.14$). $\mathcal{M}_{\text{Dvine}}$ unequivocally produces the best predictions. $\mathcal{M}_{\text{Dvine}}$ adopts streamflow data from two or three donors (Gages 3, 4 and 6) without utilizing streamflow data from the first two donors when a multiple dependence structure is established to build an ensemble prediction. It outperforms $\mathcal{M}_{\text{Bicop}}$ and $\mathcal{M}_{\text{Kraus}}$ across all validation periods, besides a few with the worst performance. Even in this case, the maximum RMSE of $\mathcal{M}_{\text{Dvine}}$ is fairly less than the maximum RMSE of $\mathcal{M}_{\text{Kraus}}$.

In addition, the ACE results present how the three models characterize prediction uncertainty. $\mathcal{M}_{\text{Dvine}}$ is capable of properly covering the predications across the entire distribution while slight overestimation occurs for the smallest two quantiles. The remaining upper quantiles also tend to slightly overestimate the true values but the overestimations are less than the other models ($\mathcal{M}_{\text{Bicop}}$, $\mathcal{M}_{\text{Kraus}}$). Taken together, the results of the synthetic experiment suggest that $\mathcal{M}_{\text{Dvine}}$ yields the best predictions among the copula-based models tested.

4.2 Performance of the copula-based models in the Yadkin-Pee Dee River

Using the insights developed from the synthetic experiment above, the three copula-based models are applied to the streamflow data for the Yadkin-Pee Dee River. At first, upper and lower tail dependences (λ_{upper} and λ_{lower}) are examined for all two pairs of sites (see Figure 4) using the approach of Schmid and Schmidt (2007). Theoretical background is described in the Supporting Information (Text S3). Note that in this analysis, the dependences become more obvious as the values approach unity. Two major insights emerge from this figure. First, many site-pairs exhibit strong upper tail dependence, suggesting that streamflow variability has a tendency to be more correlated under high-flow conditions compared to under low-flow conditions (i.e., asymmetric dependence). The lack of lower-tail dependence may be due to contributions governing low streamflow such as river regulation. Next, even under high- or low-flow conditions, there is a wide range of tail dependence across the study basin (i.e., heterogeneous dependence). To sum up, a wide range of complex dependencies is observed in the streamflow data over the study basin. The complex dependences suggest, when streamflow is estimated from multiple donors, the potential usefulness of considering a multiple dependence structure, which is one of the main features of vine copulas.

Figure 5 shows the RMSE and NSE results for the three copula-based models under a "leave-one-out" cross validation framework. This process is repeated 20 times to build an ensemble prediction by using test periods randomly defined. For this analysis, five years of data are selected to be assumed as the observed period at the target gage, and another four years are randomly selected in the remaining data for the test period. Similar to the results from the synthetic experiment, $\mathcal{M}_{\text{Kraus}}$ performs poorly in both the RMSE and NSE metrics (median

RMSE = 1.549 and NSE = 0.652). The bivariate copula performs well (median RMSE = 1.496), indicating that this approach efficiently leverages available information even though the information is limited to single donor. Particularly, $\mathcal{M}_{\text{Bicop}}$ achieves lowest RMSE values in the upper side of the RMSE box (e.g., third quartile), providing a strong uncertainty quantification skill for the upper bound. However, $\mathcal{M}_{\text{Dvine}}$ yields the best median RMSE and NSE values (= 1.359 and 0.719). Given the heterogeneous dependence conditions (see Figure 4), the high dimensional structures are effective in modeling a complex streamflow gage network. This feature can substantially improve prediction of target site flows.

Figure 6a presents the ACE scores described for principal quantiles, $\varpi \in \{0.05, 0.10, 0.20, \ldots, 0.90, 0.95\}$, across all target sites under the cross validation framework. Figure 6b presents 95% PIs for each model for an example time period (1 May 2018 to 31 July 2018) for one target site (USGS site ID: 02143500). Note that the ACE would ideally take zero value, regardless of the quantiles. The ACE scores for the three models ($\mathcal{M}_{\text{Bicop}}$, $\mathcal{M}_{\text{Kraus}}$, $\mathcal{M}_{\text{Dvine}}$) range from 0.004 to 0.0007 when considering all the quantiles together. However, the scores vary depending on the quantiles. For instance, the ACE score for $\mathcal{M}_{\text{Kraus}}$ is noticeably positive but is almost zero around the median streamflow, indicating that the model properly represent uncertainty of the median streamflow. $\mathcal{M}_{\text{Bicop}}$ and $\mathcal{M}_{\text{Dvine}}$ result in very similar ACE scores although $\mathcal{M}_{\text{Dvine}}$ performs slightly better than $\mathcal{M}_{\text{Bicop}}$. The differences in characterization of prediction uncertainty can be confirmed from a particular target site (Figure 6b).

Based on the results in Figures 5 and 6, $\mathcal{M}_{\mathrm{Dvine}}$ outperforms the other copula models (as judged by model prediction reliability and uncertainty quantification skill) and is thus selected as an appropriate copula model to infill missing data in partially gaged. Figure 7 shows an example application of $\mathcal{M}_{\mathrm{Dvine}}$ including the optimal donor sites, proper bivariate copulas and their parameters for one target site (USGS site number #214645022) when the model is calibrated using the full 15-year record.

4.3 Intermodel comparison for streamflow estimation

To assess the predictive skill of the proposed vine copula model, it is compared with six other statistical models (see Table 1). Figure 8 shows RMSE and NSE for the seven models where the streamflow values are estimated based on the available data defined by the two different cases, labeled “deficit record” and “sufficient record” (see Section 3.3). Under all cases, the vine copula approach outperforms the other infilling approaches. For example, for the “sufficient record” case, median NSE for $\mathcal{M}_{\mathrm{Dvine}}$ is 0.673 whereas those for $\mathcal{M}_{\mathrm{IDW-streamflow}}$ and $\mathcal{M}_{\mathrm{rho-streamflow}}$ are 0.462 and 0.649, respectively. In this analysis, the approaches, which are based on streamflow values of the donor sites without utilizing non-exceedance probability including DAR-streamflow and Kriging-streamflow, yield relatively increased bias in their predictions. On the other hand, an application of FDC models offers reliable predictions. For instance, for the “sufficient record” case, median RMSE for $\mathcal{M}_{\mathrm{FDC-highestrho}}$ is 1.603 compared to that of a direct of using streamflow (e.g., median RMSE of $\mathcal{M}_{\mathrm{FDC-streamflow}}$ = 3.422 for the sufficient record). Similar interpretation can be found in the comparison between $\mathcal{M}_{\mathrm{FDC-IDW}}$ and $\mathcal{M}_{\mathrm{IDW\text{-}streamflow}}$. The results from these approaches suggest that utilizing FDC process leads to a reliable estimation, which is a primary structure in the vine copula. The other noticeable feature is that available data length provides a significant influence on performance of some infilling methods. In particular, this is quite evident for the vine copula model (median RMSEs: 1.598 and 1.379 for deficit and sufficient records, respectively).

4.4 Prediction error decomposition

The RMSE is decomposed into their components (bias, variance, and timing components) for both the "deficit record" and "sufficient record" predictions (Figure 9). For the both cases, timing components primarily bring about the majority of prediction errors for all seven models. In particular, models estimating directly streamflow values (IDW-streamflow, DAR-streamflow, Kriging-streamflow) produce a somewhat biased component, which increases when a shorter record is employed in the model. For instance, the timing component for $\mathcal{M}_{\text{IDW}-\text{streamflow}}$ is 4.11 and 3.75 for the "deficit record" and "sufficient record", respectively. Moreover, timing components dominate the error metric for all cases. However, the importance of variance component is increased, especially in three models (FDC-IDW, DAR-streamflow, Kriging-streamflow). Lastly, the results inform that if the proposed vine copulas approach is adapted, variance and timing components are better captured, leading to better streamflow estimations, which is beneficial in the practical applications of water resources management.

Finally, two predictions are further produced using two additional experiments: (1) the observed marginal cumulative probabilities (i.e., using all 15 years) and conditional streamflow values constructed from the partial record (i.e., based on φ years), and (2) the estimated marginal cumulative probabilities (i.e., based on φ years) and conditional streamflow values constructed from the full record (i.e., all 15 years). Their prediction abilities are evaluated over the validated four-year period randomly selected in the remaining data. Similar to the previous analysis, each analysis is tested 20 times. The results from these experiments provide an inference to better isolate how error components from the two-step procedure (see section 2.2) influence prediction skill.

Figure 10 shows the ACE scores from the out-of-sampled predictions using the proposed Dvine model under the two scenarios. When considering all the quantiles together, the ACE scores for the two scenarios are 0.003 (scenario #1) and 0.006 (scenario #2) on average under the "deficit record" prediction. Also, the scores under the "sufficient record" prediction are all nearly 0.003. Those results of the scores are sufficiently closed to zero, implying that both predictions are reliable. Yet, compared to the predictions estimated by the cumulative probabilities estimated by the partial record, and conditional models constructed by full records (i.e., scenario #2), the ACE scores are achieved better, if the cumulative probabilities are determined by the full record, except for some of the low and high quantiles. Similar interpretation can be found in the NSE performance of two scenarios (see insets of Figure 10). It may suggest that the first procedure (i.e., how to determine the cumulative probabilities for the target site and its donors) is needed to pay careful attention when $\mathcal{M}_{\mathrm{Dvine}}$ is utilized. Nevertheless, the procedure to construct the conditional model in a streamflow gage network is obviosuly crucial since the over or under-estimations are observed in many quantiles when the insufficient sampling is employed in this process.

5. Conclusion

This study introduces a multiple dependence conditional model (i.e., vine copulas) to produce streamflow estimates at partially gaged sites. The model includes a flexible high dimensional joint dependence structure and conditional bivariate copula simulations. In order to confirm the usefulness of a multiple dependence structure and the procedure for an appropriate number of donor sites in the final vine copula model, the bivariate copula model and two types of vine copulas with their unique procedure to determine the optimal number of donor sites are first investigated using the generated data. These analyses were further extended in a case study of the Yadkin-Pee Dee River Basin, the eastern United States by estimating streamflow in partially gaged locations. In this analysis, six statistical infilling approaches were also employed to represent applicability of the proposed model.

Results of the synthetic experiment and application to the Yadkin-Pee Dee River Basin demonstrate that the propose model has benefits in some aspects. First, a multiple dependence structure adopted in the proposed model is beneficial. From the massive evaluation experiments, this study shows that multiple dependence structure clearly outperforms a single dependence structure although there is the risk of overfitting when too many dependence structures are employed. For example, the proposed model shows the improvement of 9.2 % on average compared to the bivariate model from the evaluation experiment over the historical case study. Moreover, this study confirms that the proposed multiple dependence structure model with their optimum number of donor sites produces more reliable streamflow estimation than other common infilling models. To be specific, for the “sufficient record” case, the proposed model shows the improvement of 13.9 % on average compared to the FDC-highestrho model. Next, the proposed model allows the development of confidence intervals to consider prediction uncertainty, which is fairly attractive compared to other models. For example, Bárdossy and Pegram (2013) argue that confidence intervals obtained using an ordinary kriging model do not reflect the prediction uncertainty well particularly on a daily scale. Overall, this study exhibits that a vine copula is potentially an effective tool to support water resource management planners for objectives like gap-filling or extending missing streamflow records.

While the results of the proposed model are favorable, there are possible limitations worthy of further discussion. First, the proposed method is computationally expensive, even after adopting the multicore processing to reduce the computational burden. This becomes more problematic when the method is applied to a larger, more complex streamflow gaging network. Nevertheless, because local water managers do not need to build the model repeatedly whenever they face missing values once the model is calibrated for a specific site, this computational burden may be a minor issue. Second, the assessment illustrated in this study focuses on model performance under cross-validation at partially gaged basins, but additional work is needed to extend the proposed model to ungagged basins. One possible way is to build a regression based model with spatial proximity and physical basin characteristics to define associations between the target and donor sites (e.g., Ahn and Steinschneider, 2019). Lastly, this study does not consider potential nonstationarity in FDCs and correlations caused by the influence of anthropogenic activity and change in land use. Nonstationarity may not be problematic in this analysis since the assessment is limited to 15 years across the gaging network. However, if longer records were used, it would be beneficial to consider the potential nonstationarity. This exploration is left for future work.

There are several opportunities to improve the model structure. For instance, a vine copula is able to incorporate more additional conditioning variables. One feasible approach is to add a time series of climate data (e.g., precipitation) or to decompose a time series of streamflow from the donor sites into a number of periodic components at different frequency levels through the wavelet decomposition approach (Kisi and Cimen, 2011). Moreover, although the proposed model providing a more flexible way to model multivariate dependences, it can be further improved by not adopting the standard assumption (i.e., simplifying assumption) that the conditional pair-copulas depend on the conditioning variables through the conditional margins (Acar et al., 2012). One possible alternative is the use of the semi-parametric estimation of a conditional copula (Acar et al., 2012; Vatter and Chavez-Demoulin, 2015). This semi-parametric approach enables an estimate of the dependence parameters which do not rely on the simplifying assumption, eventually leading to more reliable infilling estimations. I believe this provides an interesting avenue for future research.

Lastly, the results presented here are specific to a study basin used in a case study. The proposed model has not restricted to other watersheds around the world and its application is further required towards drawing more generalized conclusions. In addition, the model could be used for the purpose of infilling missing values of other hydrometeorological variables besides streamflow (e.g., precipitation and soil moisture). For this application, the implementation of a vine copula with combined discrete and continuous margins (i.e., to account for no rainfall days) should be explored (e.g., Stoeber et al., 2013).

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1C1C1002438). Also, the author would like to acknowledge Scott Steinschneider for his helpful comments during the development of this paper.

REFERENCES

Aas, K., Czado, C., Frigessi, A., and Bakken, H.: Pair-copula constructions of multiple dependence, Insur. Math. Econ., 44, 182–198, 2009.

Acar, E. F., Genest, C., and Nešlehová, J.: Beyond simplified pair-copula constructions, J. Multivar. Anal., 110, 74–90, 2012.

Ahn, K.-H. and Palmer, R.: Regional flood frequency analysis using spatial proximity and basin characteristics: Quantile regression vs. parameter regression technique, J. Hydrol., 540, 515–526, http://dx.doi.org/10.1016/j.jhydrol.2016.06.047, 2016a.

Ahn, K.-H. and Palmer, R. N.: Use of a nonstationary copula to predict future bivariate low flow frequency in the Connecticut river basin, Hydrol. Process., 30, 3518–3532, https://doi.org/10.1002/hyp.10876, 2016b.

Ahn, K.-H. and Steinschneider, S.: Hierarchical Bayesian Model for Streamflow Estimation at Ungauged Sites via Spatial Scaling in the Great Lakes Basin, J. Water Resour. Plan. Manag., 145, 04019030, 2019.

Aissia, M.-A. B., Chebana, F., and Ouarda, T. B.: Multivariate missing data in hydrology–Review and applications, Adv. Water Resour., 110, 299–309, 2017.

Archfield, S. A. and Vogel, R. M.: Map correlation method: Selection of a reference streamgage to estimate daily streamflow at ungaged catchments, Water Resour. Res., 46, 2010.

Ariff, N., Jemain, A., Ibrahim, K., and Wan Zin, W.: IDF relationships using bivariate copula for storm events in Peninsular Malaysia, J. Hydrol., 470, 158–171, 2012.

Arreola Hernandez, J., Hammoudeh, S., Nguyen, D. K., Al Janabi, M. A., and Reboredo, J. C.: Global financial crisis and dependence risk analysis of sector portfolios: a vine copula approach, Appl. Econ., 49, 2409–2427, 2017.

Bárdossy, A. and Pegram, G.: Interpolation of precipitation under topographic influence at different time scales, Water Resour. Res., 49, 4545–4565, 2013.

Bárdossy, A. and Pegram, G.: Infilling missing precipitation records–A comparison of a new copula-based method with other techniques, J. Hydrol., 519, 1162–1170, 2014.

Beauchamp, J., Downing, D., and Railsback, S.: Comparison of regression and time-series methods for synthesizing missing streamflow records, JAWRA J. Am. Water Resour. Assoc., 25, 961–975, 1989.

Bedford, T. and Cooke, R. M.: Probability density decomposition for conditionally dependent random variables modeled by vines, Ann. Math. Artif. Intell., 32, 245–268, 2001.

Bedford, T., Cooke, R. M., and others: Vines–a new graphical model for dependent random variables, Ann. Stat., 30, 1031–1068, 2002.

Beguería, S., Tomas-Burguera, M., Serrano-Notivoli, R., Peña-Angulo, D., Vicente-Serrano, S. M., and González-Hidalgo, J.-C.: Gap filling of monthly temperature data and its effect on climatic variability and trends, J. Clim., 32, 7797–7821, 2019.

Bevacqua, E.: CDVineCopulaConditional: Sampling from Conditional C-and D-Vine Copulas, R package version 0.1. 0, 2017.

Bhatti, M. I. and Do, H. Q.: Recent development in copula and its applications to the energy, forestry and environmental sciences, Int. J. Hydrog. Energy, 44, 19453–19473, 2019.

Blum, A. G., Archfield, S. A., and Vogel, R. M.: On the probability distribution of daily streamflow in the United States, Hydrol. Earth Syst. Sci., 21, 3093–3103, 2017.

Booker, D. and Snelder, T.: Comparing methods for estimating flow duration curves at ungauged sites, J. Hydrol., 434, 78–94, 2012.

Boscarello, L., Ravazzani, G., Cislaghi, A., and Mancini, M.: Regionalization of flow-duration curves through catchment classification with streamflow signatures and physiographic–climate indices, J. Hydrol. Eng., 21, 05015027, 2016.

Boyle, D. P., Gupta, H. V., and Sorooshian, S.: Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, Water Resour. Res., 36, 3663–3674, 2000.

Brechmann, E. C., Hendrich, K., and Czado, C.: Conditional copula simulation for systemic risk stress testing, Insur. Math. Econ., 53, 722–732, 2013.

Castellarin, A., Galeati, G., Brandimarte, L., Montanari, A., and Brath, A.: Regional flow-duration curves: reliability for ungauged basins, Adv. Water Resour., 27, 953–965, 2004.

Chen, L., Singh, V. P., Guo, S., Zhou, J., and Zhang, J.: Copula-based method for multisite monthly and daily streamflow simulation, J. Hydrol., 528, 369–384, 2015.

Croley, T. and Hartmann, H.: NOAA Technical Memorandum ERL GLERL-61: Near-Real-Time Forecasting of Large-Lake Water Supplies: A User's Manual, Ann Arbor MI, 1986.

Cunderlik, J. M. and Ouarda, T. B.: Regional flood-duration–frequency modeling in the changing environment, J. Hydrol., 318, 276–291, 2006.

Czado, C.: Pair-copula constructions of multivariate copulas, in: Copula theory and its applications, Springer, 93–109, 2010.

Czado, C.: Analyzing Dependent Data with Vine Copulas, Lect. Notes Stat. Springer, 2019.

Daneshkhah, A., Remesan, R., Chatrabgoun, O., and Holman, I. P.: Probabilistic modeling of flood characterizations with parametric and minimum information pair-copula model, J. Hydrol., 540, 469–487, 2016.

Dissmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D.: Selecting and estimating regular vine copulae and application to financial returns, Comput. Stat. Data Anal., 59, 52–69, 2013.

Erhardt, T. M., Czado, C., and Schepsmeier, U.: R-vine models for spatial time series with an application to daily mean temperature, Biometrics, 71, 323–332, 2015.

Farmer, W.: Estimating records of daily streamflow at ungaged locations in the southeast United States, PhD Disertation Tufts Univ. MA USA, 2015.

Farmer, W. H. and Vogel, R. M.: On the deterministic and stochastic use of hydrologic models, Water Resour. Res., 52, 5619–5633, 2016.

Fisk, J.: Reproductive Ecology and Habitat Use of the Robust Redhorse in the Pee Dee River, North Carolina and South Carolina., 2010.

Fu, G. and Butler, D.: Copula-based frequency analysis of overflow and flooding in urban drainage systems, J. Hydrol., 510, 49–58, 2014.

Geenens, G.: Probit transformation for kernel density estimation on the unit interval, J. Am. Stat. Assoc., 109, 346–358, 2014.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, 2009.

Hao, Z. and Singh, V. P.: Modeling multisite streamflow dependence with maximum entropy copula, Water Resour. Res., 49, 7139–7143, 2013.

He, Y., Liu, R., Li, H., Wang, S., and Lu, X.: Short-term power load probability density forecasting method using kernel-based support vector quantile regression and Copula theory, Appl. Energy, 185, 254–266, 2017.

Hughes, D. and Smakhtin, V.: Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves, Hydrol. Sci. J., 41, 851–871, 1996.

Joe, H.: Dependence modeling with copulas, CRC press, 2014.

Kalteh, A. M. and Hjorth, P.: Imputation of missing values in a precipitation–runoff process database, Hydrol. Res., 40, 420–432, 2009.

Karmakar, S. and Simonovic, S.: Bivariate flood frequency analysis. Part 2: a copula-based approach with mixed marginal distributions, J. Flood Risk Manag., 2, 32–44, 2009.

Kisi, O. and Cimen, M.: A wavelet-support vector machine conjunction model for monthly streamflow forecasting, J. Hydrol., 399, 132–140, 2011.

Kraus, D. and Czado, C.: D-vine copula based quantile regression, Comput. Stat. Data Anal., 110, 1–18, 2017.

Li, M., Shao, Q., Zhang, L., and Chiew, F. H.: A new regionalization approach and its application to predict flow duration curve in ungauged basins, J. Hydrol., 389, 137–145, 2010.

Liu, Z., Zhou, P., Chen, X., and Guan, Y.: A multivariate conditional model for streamflow prediction and spatial precipitation refinement, J. Geophys. Res. Atmospheres, 120, 10–116, 2015.

Lu, W.: A high-dimensional vine copula approach to comovement of China's financial markets, in: 2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings, 1538–1543, 2013.

Mendicino, G. and Senatore, A.: Evaluation of parametric and statistical approaches for the regionalization of flow duration curves in intermittent regimes, J. Hydrol., 480, 19–32, 2013.

Niemierko, R., Töppel, J., and Tränkler, T.: A D-vine copula quantile regression approach for the prediction of residential heating energy consumption based on historical data, Appl. Energy, 233, 691–708, 2019.

Pugliese, A., Castellarin, A., and Brath, A.: Geostatistical prediction of flow–duration curves in an index-flow framework, Hydrol. Earth Syst. Sci., 18, 3801–3816, 2014.

Salvadori, G. and De Michele, C.: Frequency analysis via copulas: Theoretical aspects and applications to hydrological events, Water Resour. Res., 40, 2004.

Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., Nagler, T., Erhardt, T., Almeida, C., Min, A., Czado, C., Hofmann, M., and others: Package 'VineCopula,' R Package Version, 2, 2015.

Schmid, F. and Schmidt, R.: Multivariate conditional versions of Spearman's rho and related measures of tail dependence, J. Multivar. Anal., 98, 1123–1140, 2007.

Schnier, S. and Cai, X.: Prediction of regional streamflow frequency using model tree ensembles, J. Hydrol., 517, 298–309, 2014.

Shafaei, M., Fakheri-Fard, A., Dinpashoh, Y., Mirabbasi, R., and De Michele, C.: Modeling flood event characteristics using D-vine structures, Theor. Appl. Climatol., 130, 713–724, 2017.

Sklar, A.: Fonctions de Répartition À N Dimensions Et Leurs Marges, Université Paris 8, 1959.

Smakhtin, V. Y.: Generation of natural daily flow time-series in regulated rivers using a non-linear spatial interpolation technique, Regul. Rivers Res. Manag. Int. J. Devoted River Res. Manag., 15, 311–323, 1999.

Stoeber, J., Joe, H., and Czado, C.: Simplified pair copula constructions—limitations and extensions, J. Multivar. Anal., 119, 101–118, 2013.

U.S. Geological Survey: National Water Information System (NWISWeb): U.S. Geological Survey database, 2018.

Vatter, T. and Chavez-Demoulin, V.: Generalized additive models for conditional dependence structures, J. Multivar. Anal., 141, 147–167, 2015.

Vernieuwe, H., Vandenberghe, S., De Baets, B., and Verhoest, N.: A continuous rainfall model based on vine copulas, Hydrol. Earth Syst. Sci., 19, 2685–2699, 2015.

Worland, S. C., Steinschneider, S., Farmer, W., Asquith, W., and Knight, R.: Copula theory as a generalized framework for flow-duration curve based streamflow estimates in ungaged and partially gaged catchments, Water Resour. Res., 55, 9378–9397, 2019.

Xu, D., Wei, Q., Elsayed, E. A., Chen, Y., and Kang, R.: Multivariate degradation modeling of smart electricity meter with multiple performance characteristics via vine copulas, Qual. Reliab. Eng. Int., 33, 803–821, 2017.

Xu, Q. and Childs, T.: Evaluating forecast performances of the quantile autoregression models in the present global crisis in international equity markets, Appl. Financ. Econ., 23, 105–117, 2013.

Zaman, M. A., Rahman, A., and Haddad, K.: Regional flood frequency analysis in arid regions: A case study for Australia, J. Hydrol., 475, 74–83, 2012.

Zimmer, D. M.: Analyzing comovements in housing prices using vine copulas, Econ. Inq., 53, 1156–1169, 2015.

List of Figures

List of Tables

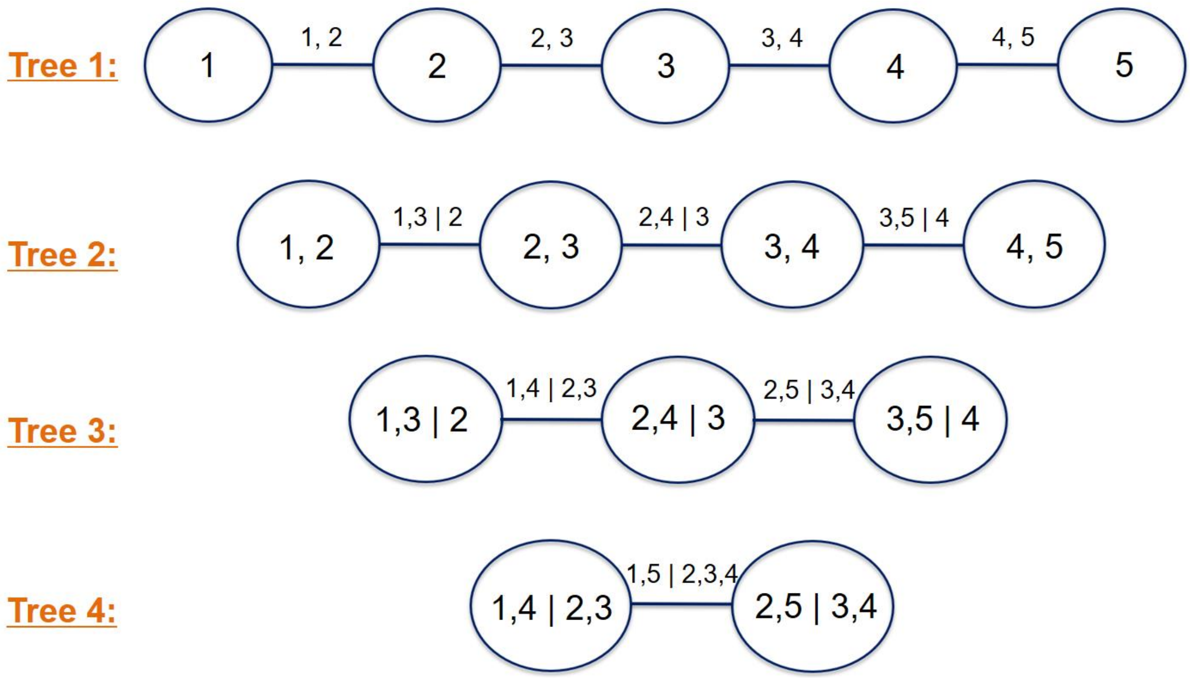

Figure 1 Example of D-vine structures with 5 variables, 4 trees and 10 edges

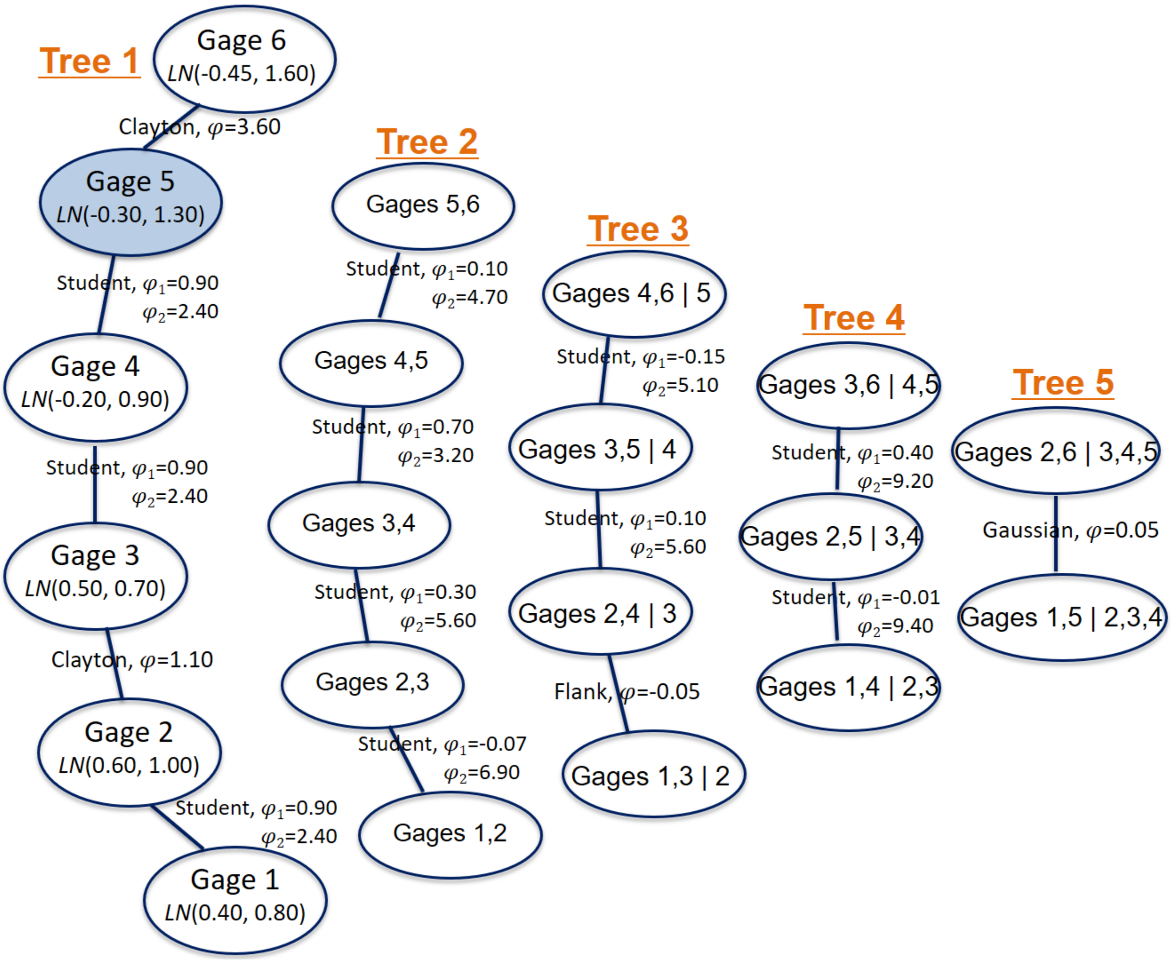

Figure 2 Structure of the 6-Dimensional vine model and marginal for the synthetic simulation. $LN(\pi, \sigma^2)$ denotes the log normal distribution with its mean (π) and variance (σ^2). The target gage is highlighted.

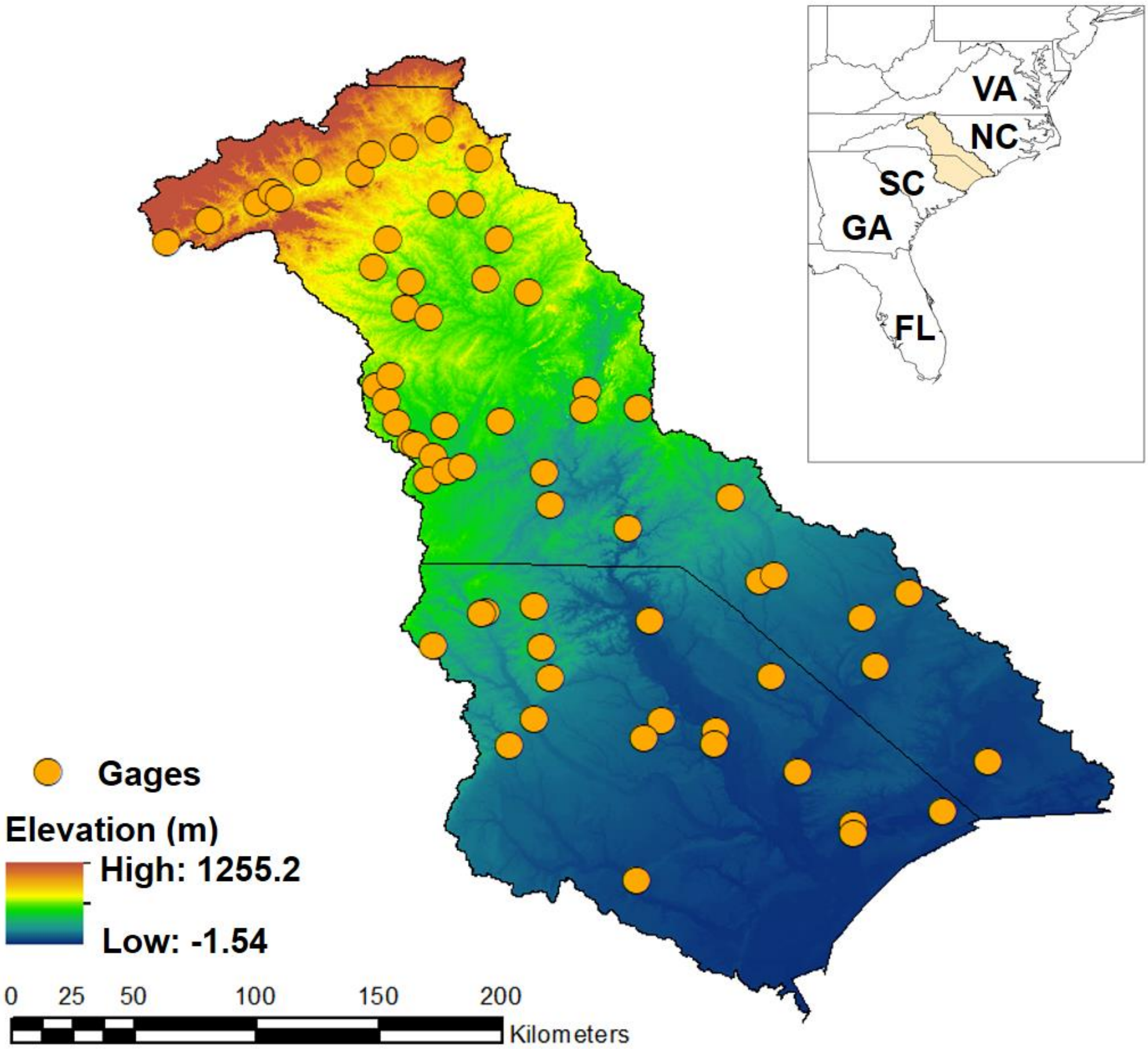

Figure 3 Map of the Yadkin-Pee Dee Basin with 54 stream gage stations

.

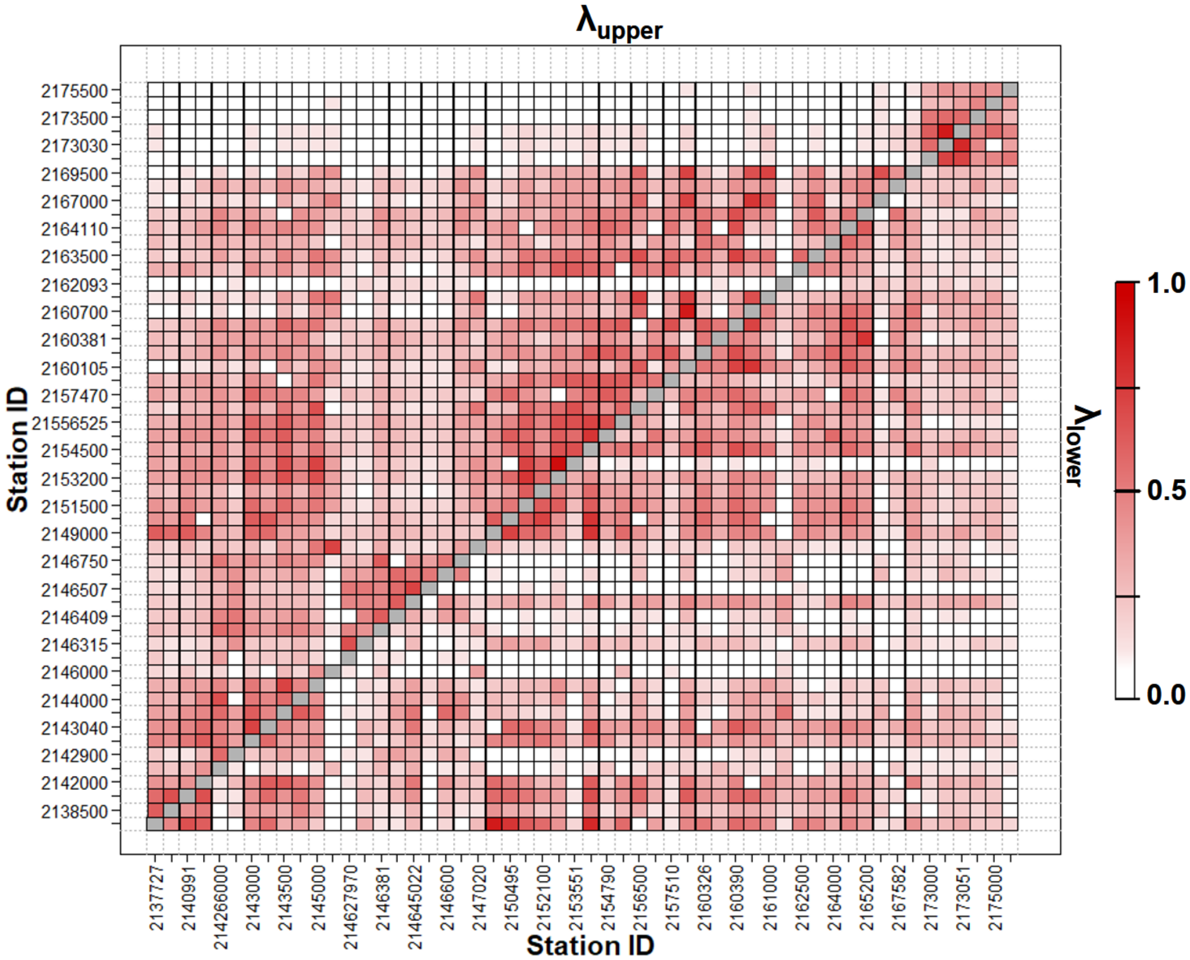

Figure 4 Pairwise upper and lower tail dependence for watersheds in the Yadkin-Pee Dee River Basin. The upper triangular matrix shows values for the upper-tail dependence and the lower triangular matrix presents values for the lower-tail dependence. The metrics can range from 0 to 1, with higher values suggesting greater interdependence of two streamflows for each upper- and lower-tail.

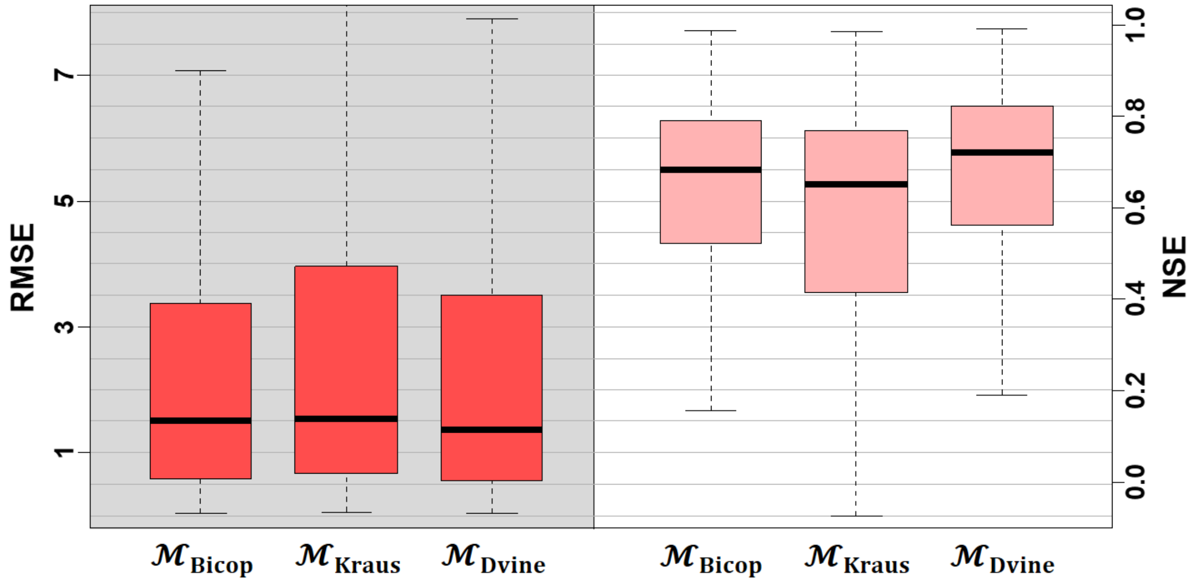

Figure 5 Model performance for the Yadkin-Pee Dee river under a cross-validation framework based on RMSE (dark squares) and NSE (light squares). Here, the RMSE (NSE) can range from 0 to ∞ ($-\infty$ to 1), with lower RMSE (higher NSE) implying better performance.

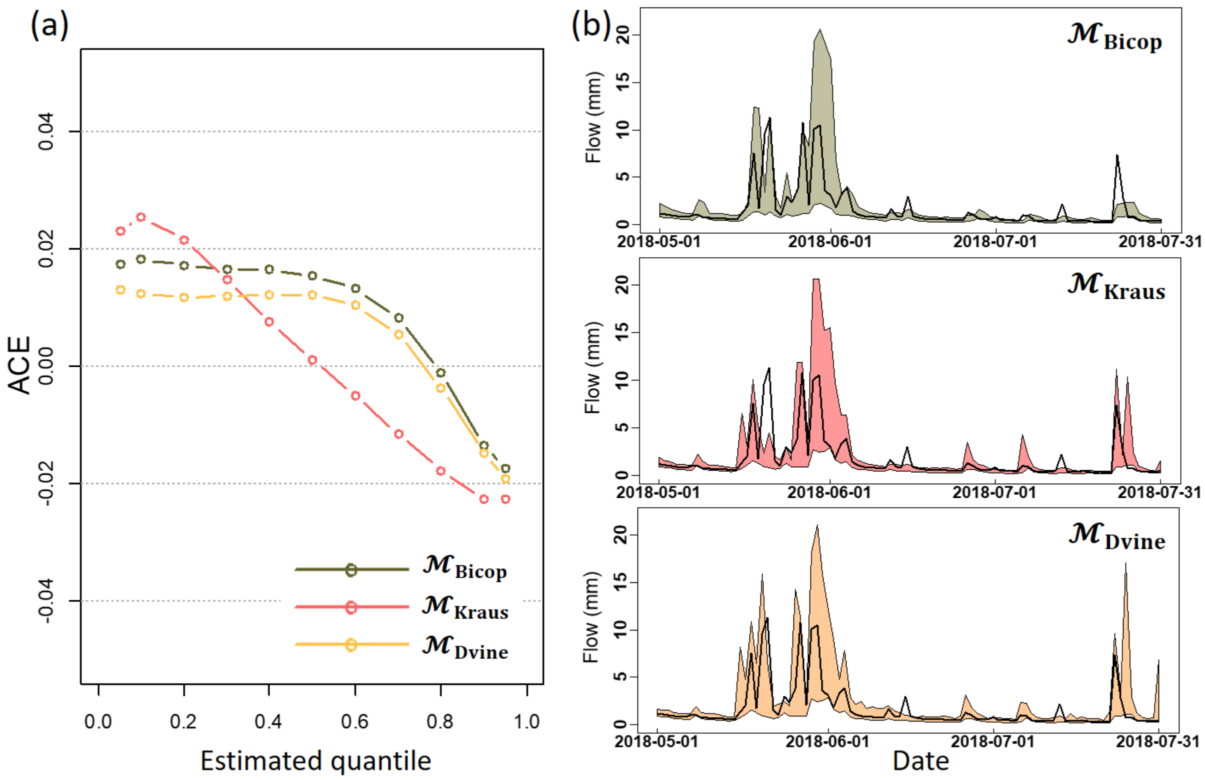

Figure 6 (a) Average coverage error from three copula-based models for the Yadkin-Pee Dee River Basin across exemplarily quantiles, and (b) 95% PIs for three models for an example period (1 May 2018 to 31 July 2018) for a specific target gauge (USGS site ID: 02143500). Observed streamflow (black solid line) is also presented in each figure.

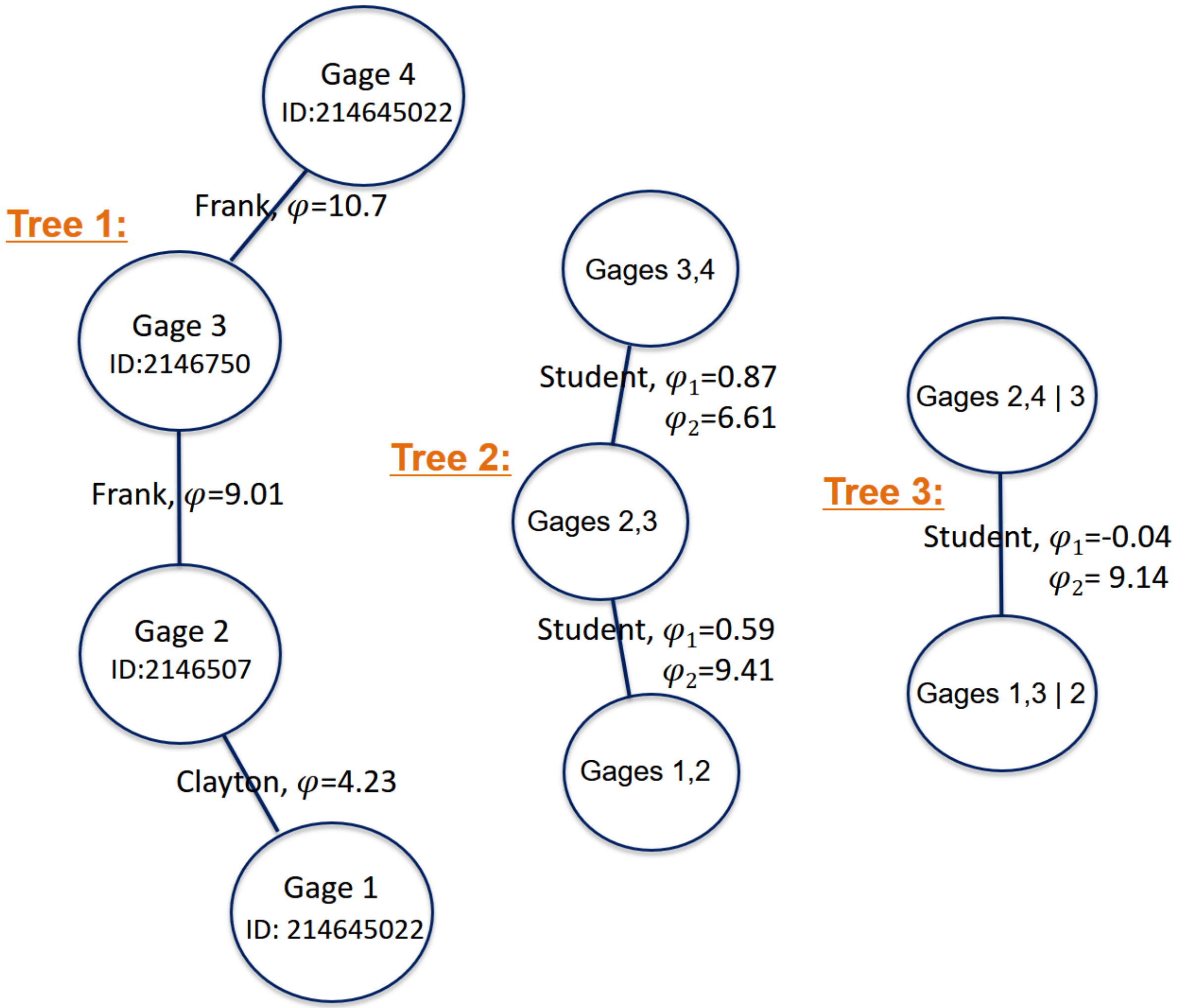

Figure 7 Structure of the Dvine copula applied for a particular target site (USGS site ID: 214645022) with the defined bivariate copulas and their parameters.

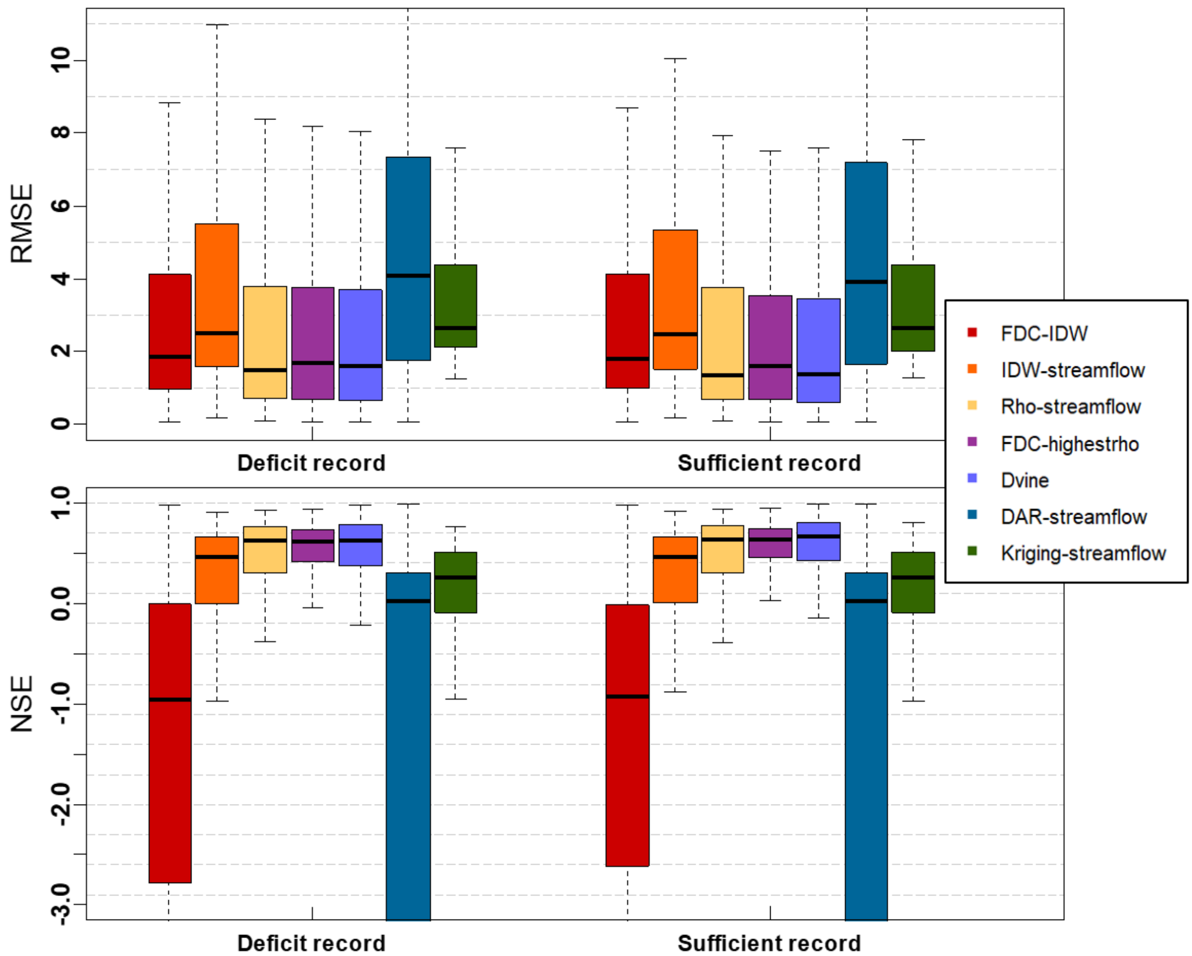

Figure 8 Inter-model comparison using cross-validation experiments based on RMSE (upper) and NSE (lower). Here, lower RMSE suggests more accurate estimations for infilling missing values.

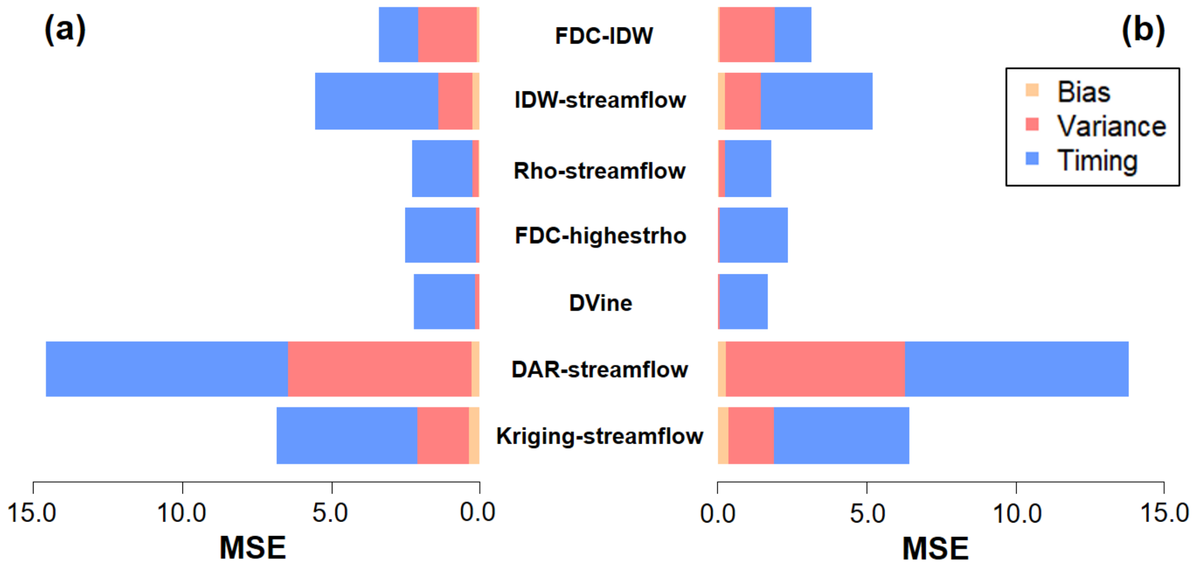

Figure 9 Three contributions from the decomposed mean squared error (MSE) for the cross-validation experiment with (a) the deficit record and (b) sufficient record scenarios.

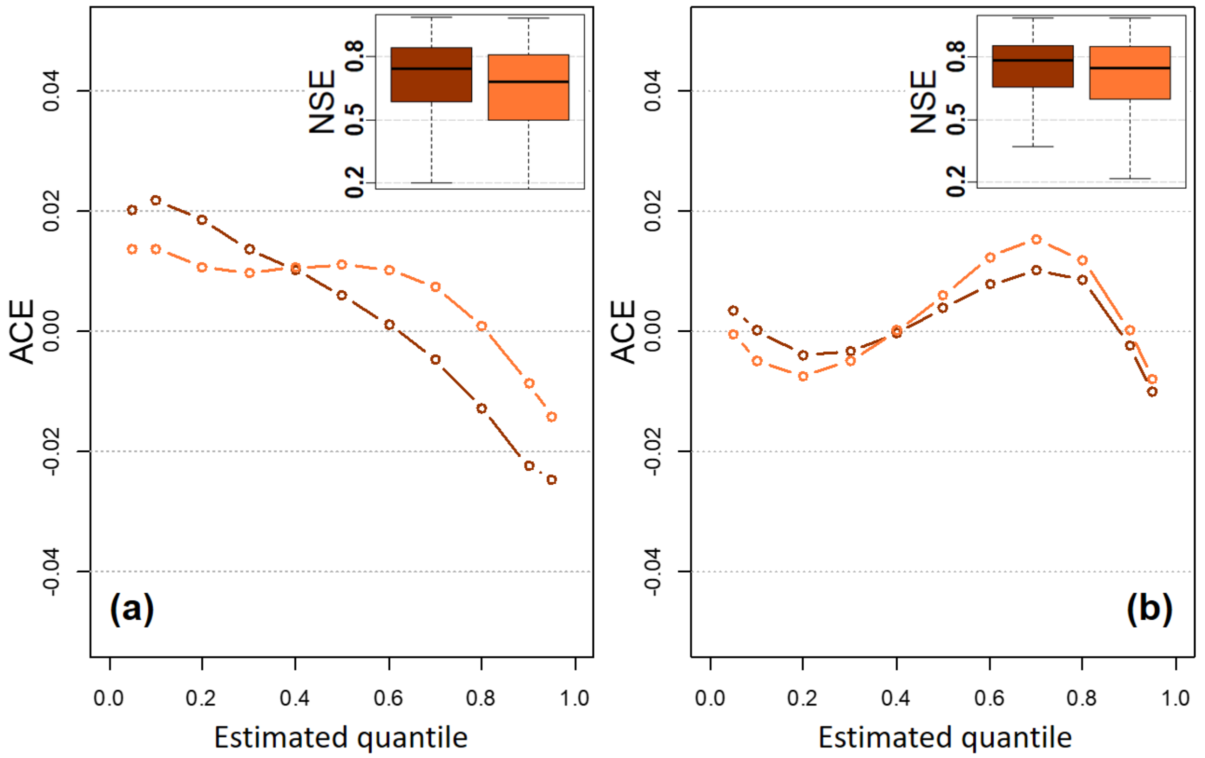

Figure 10 Average coverage error of the Dvine model for two scenarios under (a) the "deficit" and (b) "sufficient" cases. In each case, the dark line represents the scenario by the marginal cumulative probabilities using all years and conditional streamflow values constructed from the partial record. On the other hand, the light line illustrates the scenario by the marginal cumulative probabilities estimated by the partial record and conditional streamflow values constructed from the full record. Inset: NSE performance of the Dvine model for the two scenarios in each case.

Table 1 Seven infilling approaches discussed in the study

No.	Method	Description
1	FDC-IDW	Inverse distance-weighted estimate of non-exceedance probability from those of all donors.
2	IDW-streamflow	Inverse distance-weighted estimate using streamflow from all donors.
3	Rho-streamflow	Correlation-weighted streamflow estimate from the selected donors for each time step. The optimal number of donors is determined in a cross-validation framework.
4	FDC-highestrho	Estimate non-exceedance probability from the gage with the highest correlation.
5	DAR-streamflow	Drainage-area (DA) ratio for streamflow using the DA from the nearest neighbor gage.
6	Kriging-streamflow	Geostatistical interpolation method to estimate streamflow from all donors for each time step.
7	DVine	Vine copula-based estimate from the selected donors

Table 2 RMSE and NSE results over the validation periods under synthetic experiment for comparing copula-based model formulations. Best metric values for each quantile are italicized and bolded.

Metric	Model formulation	Min	First quantile	Median	Third quantile	Max
Root mean squared error (RMSE)	$\mathcal{M}_{\mathrm{Bicop}}$	0.912	1.119	1.258	1.363	***3.353***
	$\mathcal{M}_{\mathrm{Kraus}}$	0.990	1.140	1.386	1.660	4.273
	$\mathcal{M}_{\mathrm{Dvine}}$	***0.895***	***1.046***	***1.112***	***1.391***	4.119
Nash-Sutcliffe efficiency (NSE)	$\mathcal{M}_{\mathrm{Bicop}}$	***0.464***	0.779	0.826	0.856	0.902
	$\mathcal{M}_{\mathrm{Kraus}}$	0.198	0.724	0.782	0.825	0.885
	$\mathcal{M}_{\mathrm{Dvine}}$	0.248	***0.805***	***0.838***	***0.869***	***0.905***

Table 3 Results of average coverage error (ACE) over the validation periods under synthetic experiment for comparing copula-based model formulations. Best metric values for each quantile are italicized and bolded.

Model formulation	Estimated quantile (ϖ)				
	0.05	0.10	0.50	0.90	0.95
$\mathcal{M}_{\mathrm{Bicop}}$	0.027	0.063	0.079	0.014	0.002
$\mathcal{M}_{\mathrm{Kraus}}$	***0.003***	***0.011***	0.055	0.024	0.001
$\mathcal{M}_{\mathrm{Dvine}}$	0.029	0.048	***0.042***	***0.001***	***0.000***