Comment on “A coupled stochastic rainfall-evapotranspiration model for hydrological impact analysis” by Minh Tu Pham et al.

Thomas Nagler

June 16, 2017

1 General comments

The manuscript is very well written and gives sufficient context to understand the relevant developments and issues in hydrological impact analysis. The authors clearly motivate why a stochastic rainfall-evapotranspiration model is useful in this context. Their proposal is based on vine copulas, a modern statistical tool for modeling stochastic dependence between multiple variables. This is a laudable effort, but the way this methodology is applied and its performance is evaluated is problematic in several ways. I fully acknowledge that HESS is not a statistics journal and statistical subtleties may not matter in specific applications. But the extent to which they do in this particular context are unclear and needs to be addressed.

Below I identify three major issues and explain why they are problematic from a statistical perspective. I urge the authors to thoroughly evaluate the implications for their hydrological model. Where possible, I try to make suggestions for alternative methods that may improve their model and its assessment. Since the first two issues may be equally relevant for other readers, my comments will be more elaborate than what is common in a closed review.

2 Specific comments

2.1 Major issues

2.1.1 Seasonal effects

A copula models the dependence between two random variables X_1, X_2 with marginal distributions F_1 and F_2. Its parameters can be estimated from observations of these variables, $\boldsymbol{X}_t = (X_{1,t}, X_{1,t})$, $t = 1, \dots, T$. The usual assumption for the validity of the estimate is that the data $\boldsymbol{X}_t$ are independent and identically distributed (*iid*). In particular, the distribution of $\boldsymbol{X}_t$ should not change with t, which is usually violated by climatic variables. The authors acknowledge that by fitting multiple models, one for each month.

I am afraid this may not be enough, because climatic trends also exist within months. For example: in central Europe, the end of April is — on average — much warmer than the beginning of April. Suppose that additionally, the average precipitation is decreasing during April. Then high temperatures will likely coincide with low levels of precipitation and vice versa. A copula fitted to this data will show negative dependence, which merely reflects the two deterministic within-month trends working in opposite directions, but not the stochastic dependence between the time series.

Whether or not within-month trends exist can be easily checked visually or by formal statistical tests (e.g., Harris and Sollis, 2003, Chapter 3). If they do exist, they should be accounted for on a finer time scale. Splitting the data into weeks or even days could be a solution, but significantly decreases the number of observations available for fitting the copula model. A good alternative is to center and scale the time series by their seasonal mean and standard deviation. More specifically, if $X_{j,d,y}$ denotes variable j observed at day d of year y, set

$$\tilde{X}_{j,d,y} = (X_{j,d,y} - \mu_{j,d})/\sigma_{j,d}, \tag{1}$$

where $\mu_{j,d}$ and $\sigma_{j,d}$ are the mean and standard deviation of $X_{j,d,y}$, $y = 1, \dots, 72$. If necessary, trends in the skewness of $\tilde{X}_{j,d,y}$ can be removed similarly using a Box-Cox transformation. This transformation is usually sufficient to account for deterministic seasonal effects. We can now build a copula for the stochastic dependence in $\tilde{\boldsymbol{X}}_t = \tilde{\boldsymbol{X}}_{d+365(y-1)}$. Simulated data from this model can be transformed to the original scale by inverting (1).

2.1.2 Inter-serial dependence

Even when the distribution of $\tilde{X}_t$ is the same for each t, subsequent observations of the time series may not be independent. Such data is called *stationary* which is less restrictive than *iid*. Typically, stationarity is sufficient to allow for valid estimation of the marginal distributions and copula of $\tilde{X}_t$. But inference tools (like confidence intervals and goodness-of-fit tests) derived under the *iid* assumption are no longer valid.

Another potential issue is that inter-serial dynamics can play an important role in applications. If so, these dynamics should be modeled explicitly explicitly. In the context of hydrological discharges, this is likely the case. Large discharges often occur when extreme weather conditions have been persistent for several days, and persistence is a sign of inter-serial dependence. A simple way to check whether such dependence is present is to look at the autocorrelation of the time-series, i.e., the correlation between $\tilde{X}_t$ and $\tilde{X}_{t-1}$ (and their squares). If the correlation is small, one can test statistically whether it is zero.

If there is dependence, there are two popular ways to capture it:

1. **Copula models**: This route is taken by the authors in 2.3.3, but only for the temperature variable. Similar models for the inter-serial dependence in evapotranspiration and precipitation should be employed in addition. If $F_{j,t,t-1}$ is the joint distribution of $\tilde{X}_{j,t}$ and $\tilde{X}_{j,t-1}$, the between-variables dependence can be modeled by a copula for the variables

$$U_{1,t} = F_{1,t|t-1}(\tilde{X}_{1,t} \mid \tilde{X}_{1,t-1}), \quad U_{2,t} = F_{2,t|t-1}(\tilde{X}_{2,t} \mid \tilde{X}_{2,t-1}),$$

where $F_{j,t|t-1}$ is the conditional distribution of $\tilde{X}_{j,t}$ given $\tilde{X}_{j,t-1}$.

2. **Classical time series models**: Classical time series models (see, e.g., Shumway et al., 2000, Chapter 3) assume that the variable $\tilde{X}_{j,t}$ is a linear combination of the preceding values ($t' < t$) and *iid* noise. For example, the autoregressive model of order p is

$$\tilde{X}_{j,t} = \sum_{k=1}^{p} \phi_{j,k} \tilde{X}_{j,t-k} + \epsilon_{j,t},$$

where $\phi_{j,k}$ are model parameters and $\epsilon_{j,t}$ is *iid* noise with mean zero. The sequence $\epsilon_{j,t}$ is commonly called *innovation* or *residual* series. The stochastic between-variable dependence can then be captured by a copula model for $(\epsilon_{1,t}, \epsilon_{2,t})$. More complex models are required when $\tilde{X}_{j,t}^2$ is autocorrelated (Harris and Sollis, 2003, Chatper 8).

2.1.3 Assessing the quality of the vine copula model

There are multiple issues with how the quality of the model is evaluated:

1. To check the model's validity, the authors merely look at the density/cdf of the observed and simulated values of a single time series. This is only weakly related to the vine copula model and not a good indicator for its fit. Under this measure, just simulating from the distribution F_E (thereby assuming that E is independent of T and P) would lead to results that are at least as good as the ones from the vine copula.

   To adequately assess the quality of the dependence model, pair-wise comparisons should be made. For example, one can look at the scatter plots of observed and simulated pairs $(X_{1,t}, X_{2,t})$. Another alternative are contour plots of the estimated joint density of observed vs. simulated pairs. Such comparisons should be made for all variable combinations. Similarly, multivariate return periods should be considered instead of single-variable return periods (see, Salvadori et al., 2011).

2. Figures 6 and 9 use empirical cumulative distribution functions (ECDF) instead of densities for no obvious reason. I advise against using ECDF's because they suggest a misleading sense of closeness between distributions. Since ECDFs are necessarily monotone functions with boundary values 0 and 1, their shape is quite restricted. For example, the left panels of Figure 9(d) show that the distributions are different, but the ECDFS still look somewhat similar. But the corresponding densities would show almost no overlap and more clearly communicate the dissimilarity.

3. In Section 4, the uncertainty in the simulation model is assessed for various degrees of data availability. From the spread of estimated densities in Figures 11-15, the authors conclude that uncertainty increases when a variable is not observed and needs to be simulated. This is likely true, but can not be inferred from these figures. The density plots for cases 1-3 are based on a different number of observations. The spread seen on 125000 simulations will naturally be larger than the the spread on 50 observations — even when the actual distribution is the same. Hence, the spreads should only be compared when they are based on the same number of simulations.

2.2 Minor issues

1. Since vine copulas are the essential ingredient in your model, I suggest to indicate this in the abstract.

2. p. 4, p. 165: If unconditional bivariate copulas are used (as is common), a vine copula is not a decomposition, but a construction. A decomposition is called *non-simplified vine copula* and involves conditional bivariate copulas (see, e.g., Stöber et al., 2013). I suggest to rephrase this sentence.

3. p. 4, l. 167: I suggest to change "all types of dependence" to "a wide range of dependence structures". "All types" can only be modeled by a non-simplified vine copula.

4. p. 5, l. 179: What do you mean by "C-vine copulas are easier to construct than D-vine copulas"? In fact, any three-dimensional vine is both a C- and D-vine, which can be easily verified by re-arranging the vertices of the vine graph.

5. p. 6, l. 203 ff.: I am afraid a reader without prior knowledge of vine copulas will not understand your paragraph on how the model is estimated. Instead of your explanation, it should suffice to refer the reader to Aas et al. (2009).

6. How are marginal distributions modeled/estimated?

7. Figures 4, 12–17 should use a larger smoothing parameter to decrease variability of the density estimates. A large proportion of the observed variability is due to the density estimation technique. This is not the kind of variability you want to assess.

8. Figure 22: What is i?

3 Technical corrections

1. p. 5, l. 179: "Sine because C-vine . . . " should be "Because C-vine . . . ".

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2):182–198.

Harris, R. and Sollis, R. (2003). *Applied time series modelling and forecasting*. Wiley.

Salvadori, G., De Michele, C., and Durante, F. (2011). On the return period and design in a multivariate framework. *Hydrology and Earth System Sciences*, 15(11):3293–3305.

Shumway, R. H., Stoffer, D. S., and Stoffer, D. S. (2000). *Time series analysis and its applications*, volume 3. Springer.

Stöber, J., Joe, H., and Czado, C. (2013). Simplified pair copula constructions—limitations and extensions. *Journal of Multivariate Analysis*, 119:101–118.