Hierarchical Vine copula models for the analysis of glacier discharge

Author: Mario Gómez Díaz.

Advisors: M. Concepción Ausín Olivera. Universidad Carlos III.

M. Carmen Domínguez Álvarez. Universidad de Salamanca.

Nov 8th, 2017

Index

Index

Glacier discharge

"The study of the mass balance in glaciers is crucial for the accurate quantification of water resources" (Hamlet and Lettenmaier, 1999)

"As consequence of Global Warming the ice cap glaciers lose mass through the so-called processes of ablation" (Hock et al., 2005)

- Solid loss: Calving icebergs.
- Liquid loss: Through surface melting and run off, or melting and exit at the base.

This liquid loss is known as glacier discharge.

Glacier discharge

Classification of glacier models

- Energy balance models: Complex equations.
- Degree-Day models: Assume linearity.

Glacier discharge

Classification of glacier models

- Energy balance models: Complex equations.
- Degree-Day models: Assume linearity.

Our proposal: Use copulas to model the relationship between variables

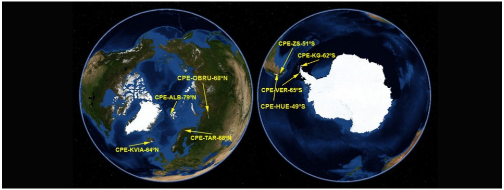
GLACKMA

GLACKMA ▸ www.glackma.es studies the glaciers as sensors of the Global Warming.

They use their studies to make public the reality of the polar regions.

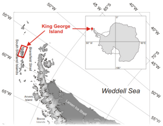
GLACKMA has got a network of 8 stations that register the glacier discharge since 2001.

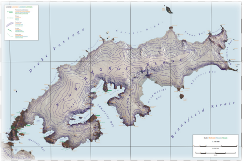
GLACKMA

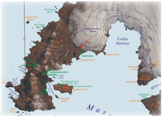


We have focused on the data from the station CPE-KG-62°, sited in King George island, that collect data from Collins glacier.

"Antarctic Peninsula is one of the most sensitive areas to the Global Warming" (Rückamp et al. 2011).







Data base

The available discharge data for this project are from 01/10/2002 to 30/09/2012.

- Specific Glacier Discharge ($m^3/sec \cdot km^2$), (Domínguez, 2004).
- Air Temperature ($^\circ C$)
- Solar Radiation ($watt/m^2$)
- Humidity (%)
- Accumulated Precipitation (mm)

Data base

The available discharge data for this project are from 01/10/2002 to 30/09/2012.

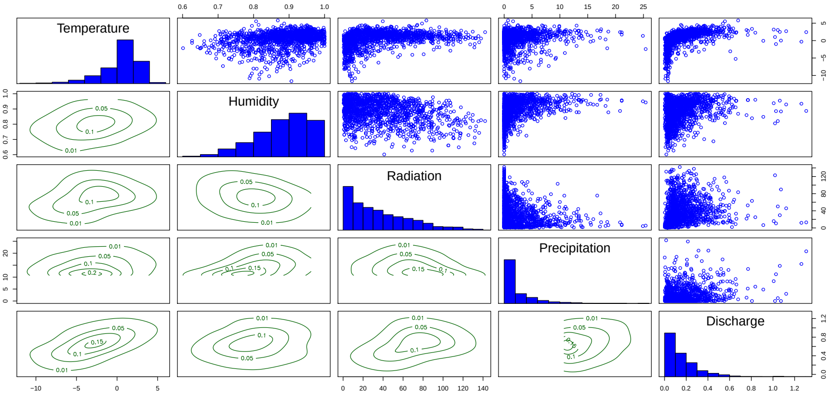
- Specific Glacier Discharge ($m^3/sec \cdot km^2$), (Domínguez, 2004).
- Air Temperature ($^\circ C$)
- Solar Radiation ($watt/m^2$)
- Humidity (%)
- Accumulated Precipitation (mm)

Large amount of zero values

- 62% of the discharge observations
- 31% of the precipitation observations

Relationship among the variables

There is relation among the variables but, apparently, is not linear.



Copulas

Definition: "Copulas are statistical instruments that allow us modelling the relationship among the variables independently of the marginal distributions choice"

Joint Distribution

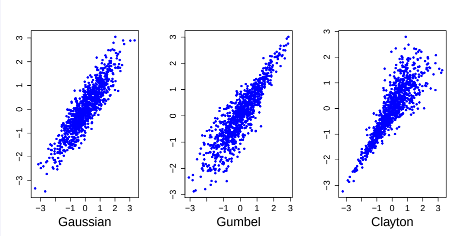
$$F(x_1, x_2, \ldots, x_n) = C\left(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)\right)$$

Density function

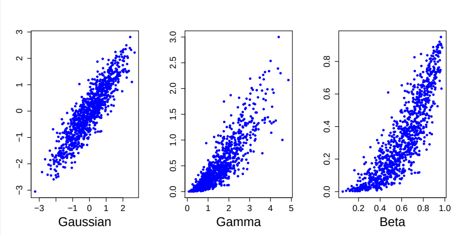
$$f(x_1, x_2, \ldots, x_n) = c\left(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)\right) \cdot f_1(x_1) \cdot f_2(x_2) \ldots \cdot f_n(x_n)$$

Bivariate copulas

Same marginal - Different copulas



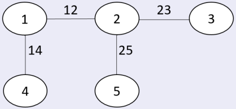
Different marginals - Same copula



Vine structure

Bedford and Cooke (2001) introduced a graphical structure called regular vine structure

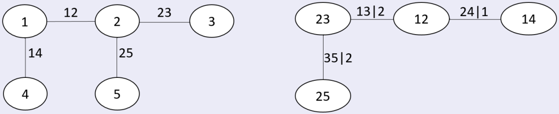
Regular vine structure



Vine structure

Bedford and Cooke (2001) introduced a graphical structure called regular vine structure

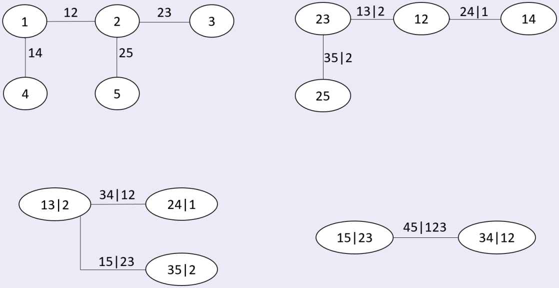
Regular vine structure



Vine structure

Bedford and Cooke (2001) introduced a graphical structure called regular vine structure

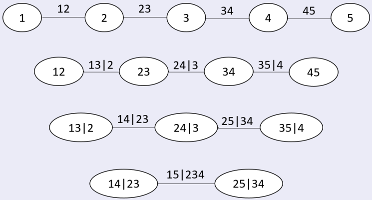
Regular vine structure



vine copulas

There are two special regular vine structures

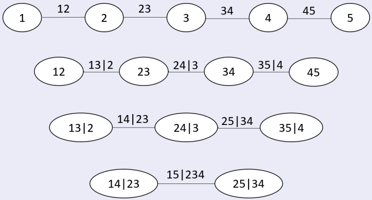
d-vine: Every tree is a path



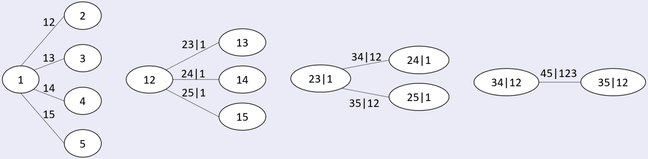
vine copulas

There are two special regular vine structures

d-vine: Every tree is a path



c-vine: Every tree has a node that is joint with the others



Index

Contribution

- Multivariate Copula Model includes all the variables.
- Use c-vine structures to model the relationship.
- Explores the seasonality through the division into periods for,
  - The marginal parameters.
  - The relationship between the variables.
- Solves the "zero-problem" through the division of the data into groups.

Joint distribution

Now we model the relationship among temperature, humidity, radiation, precipitation and discharge.

There are days with discharge zero and days with precipitation zero, thus the joint density function can be written as,

$$f(t,h,r,p,d) = \begin{cases} f_{thr}^{00}(t,h,r), & \text{with } \Pr(D=0,P=0) \\ f_{thrd}^{10}(t,h,r,d), & \text{with } \Pr(D>0,P=0) \\ f_{thrp}^{01}(t,h,r,p), & \text{with } \Pr(D=0,P>0) \\ f_{thrpd}^{11}(t,h,r,p,d), & \text{with } \Pr(D>0,P>0) \end{cases}$$

Each one of these joint distributions will be expressed with a c-vine copula

Joint distribution

Joint distribution

$$
\begin{aligned}
f^{11}_{thrpd}(t,h,r,p,d) &= f(d \mid t,h,r,p)\cdot f(t,h,r,p) \\
&= f(d \mid t,h,r,p)\cdot f(p \mid t,h,r)\cdot f(t,h,r) \\
&= f(d \mid t,h,r,p)\cdot f(p \mid t,h,r)\cdot f(r \mid t,h)\cdot f(t,h) \\
&= f(d \mid t,h,r,p)\cdot f(p \mid t,h,r)\cdot f(r \mid t,h)\cdot f(h \mid t)\cdot f(t),
\end{aligned}
$$

Joint distribution

Joint distribution

$$
\begin{aligned}
f^{11}_{thrpd}(t,h,r,p,d) &= f(d \mid t,h,r,p)\cdot f(t,h,r,p)\\
&= f(d \mid t,h,r,p)\cdot f(p \mid t,h,r)\cdot f(t,h,r)\\
&= f(d \mid t,h,r,p)\cdot f(p \mid t,h,r)\cdot f(r \mid t,h)\cdot f(t,h)\\
&= f(d \mid t,h,r,p)\cdot f(p \mid t,h,r)\cdot f(r \mid t,h)\cdot f(h \mid t)\cdot f(t),
\end{aligned}
$$

Decomposition with the c-vine structure

$$
\begin{aligned}
f^{11}_{thrpd}(t,h,r,p,d) = {} & f^{11}_t(t)\cdot f^{11}_h(h)\cdot f^{11}_r(r)\cdot f^{11}_p(p)\cdot f^{11}_d(d)\\
&\cdot c_{th}(u^{11}_t, u^{11}_h)\cdot c_{tr}(u^{11}_t, u^{11}_r)\cdot c_{tp}(u^{11}_t, u^{11}_p)\cdot c_{td}(u^{11}_t, u^{11}_d)\\
&\cdot c_{hr|t}(u^{11}_{h|t}, u^{11}_{r|t})\cdot c_{hp|t}(u^{11}_{h|t}, u^{11}_{p|t})\cdot c_{hd|t}(u^{11}_{h|t}, u^{11}_{d|t})\\
&\cdot c_{rp|th}(u^{11}_{r|th}, u^{11}_{p|th})\cdot c_{rd|th}(u^{11}_{r|th}, u^{11}_{d|th})\cdot c_{pd|thr}(u^{11}_{d|thr}, u^{11}_{p|thr}),
\end{aligned}
$$

Joint distribution

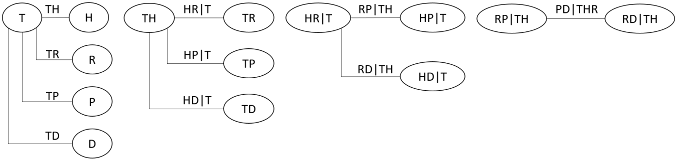
$$u_x^{11} = F_x(x), \quad x = t, h, r, p, d$$

$$u_{x|t}^{11} = \frac{\partial C_{tx}(u_t^{11}, u_x^{11})}{\partial u_t^{11}}, \quad x = h, r, p, d$$

$$u_{x|th}^{11} = \frac{\partial C_{hx|t}(u_{h|t}^{11}, u_{x|t}^{11})}{\partial u_{h|t}^{11}}, \quad x = r, p, d$$

$$u_{x|thr}^{11} = \frac{\partial C_{rx|th}(u_{r|th}^{11}, u_{x|th}^{11})}{\partial u_{r|th}^{11}}, \quad x = p, d$$

5-NODE C-VINE COPULA



Marginal distributions

For example, the density function of the temperature can be expressed as:

$$f_t(t) = \begin{cases} f_t^{00}(t) = f_t(t \mid D = 0, P = 0), \text{ with } \Pr(D = 0, P = 0) \\ f_t^{10}(t) = f_t(t \mid D > 0, P = 0), \text{ with } \Pr(D > 0, P = 0) \\ f_t^{01}(t) = f_t(t \mid D = 0, P > 0), \text{ with } \Pr(D = 0, P > 0) \\ f_t^{11}(t) = f_t(t \mid D > 0, P > 0), \text{ with } \Pr(D > 0, P > 0). \end{cases}$$

Finite mixture of distributions for every variable

- Gaussian for temperature
- Beta for humidity
- Gamma for radiation, precipitation and discharge

Conditional Probability

We can determine the probability of discharge given the value of the meteorological variables (T, H, R, P):

Conditional probability

$$P(D = 0 \mid T = t, H = h, R = r, P = p) = \frac{f(t, h, r, p \mid D = 0) \cdot \Pr(D = 0)}{f(t, h, r, p)},$$

Conditional Probability

We can determine the probability of discharge given the value of the meteorological variables (T, H, R, P):

Conditional probability

$$P(D = 0 \mid T = t, H = h, R = r, P = p) = \frac{f(t, h, r, p \mid D = 0) \cdot \Pr(D = 0)}{f(t, h, r, p)},$$

Or obtain the conditional distribution function of the discharge given the value of the meteorological variables:

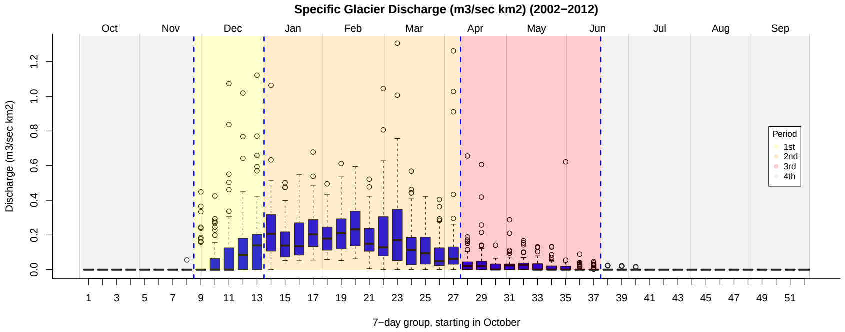
Conditional distribution

$$F_D(d \mid T = t, H = h, R = r, P = p) = \begin{cases} 0, & \Pr(D = 0 \mid t, h, r, p) \\ F(d \mid t, h, r, p, D > 0), & 1 - \Pr(D = 0 \mid t, h, r, p) \end{cases}$$

Seasonality

We divide each hydrological year in four periods to capture the seasonality.

Period	Start	Description
1	26th November	Discharge start period. Days can have positive or zero discharge.
2	31st December	Main discharge period. Almost every day have positive discharge.
3	8th April	Discharge end period. Days can be zero or positive discharge.
4	16th June	Zero discharge period. There is always zero discharge.

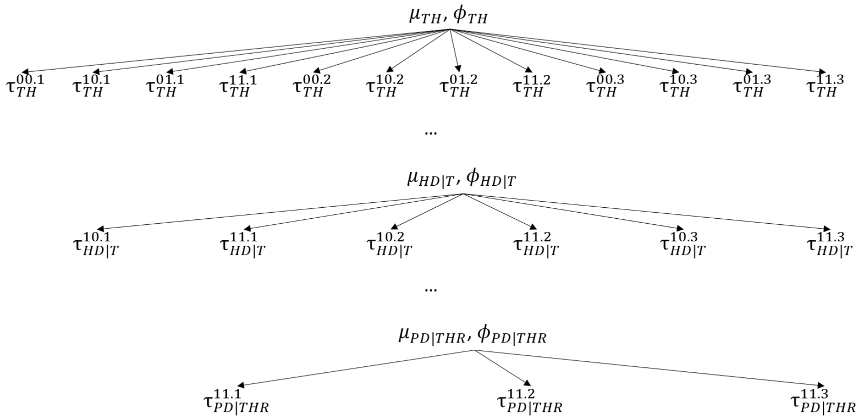


Index

Contribution

- Hierarchical Copula Model over groups and periods.
- IFM (Joe 1997). First marginal parameters, then copula parameters and hiperparameters
- ABC algorithm to make inference.
- Comparison between RWMH and ABC algorithms.

Hierarchical model



Hierarchical Model

$$\phi_j \sim \mathcal{G}\left(\frac{a_j}{2}, \frac{b_j}{2}\right),$$

$$\mu_j \mid \phi_j \sim \mathcal{N}\left(m_j, \frac{1}{\alpha_j \cdot \phi_j}\right),$$

$$\xi_j^k \mid \mu_j, \phi_j \sim \mathcal{N}\left(\mu_j, \frac{1}{\phi_j}\right),$$

$$\tau_j^k = h\left(\xi_j^k\right) = 2 \cdot logit^{-1}\left(\xi_j^k\right) - 1, \tau \in (-1, 1),$$

$$\tau_j^k = h\left(\xi_j^k\right) = logit^{-1}(\xi_j^k), \tau \in (0, 1)^{(*)},$$

$$\theta_j^k = g\left(\tau_j^k\right),$$

$^{(*)}\tau \in (0, 1)$ for Gumbel, Clayton y Joe

RWMH

1. Sample $\mu_j, \phi_j \mid \boldsymbol{\tau_j}$
2. Sample $\boldsymbol{\tau_j} \mid \boldsymbol{u_j}, \mu_j, \phi_j$

For the first step we have a Normal-Gamma distribution, then we know the conjugate posterior distribution and we can sample directly from it:

$$\begin{aligned}
\mu_j, \phi_j \mid \boldsymbol{\tau}_j &\sim \mathcal{NG}\left(m_j^*, \alpha_j^*, \frac{a_j^*}{2}, \frac{b_j^*}{2}\right), \text{where} \\
\boldsymbol{\xi}_j &= h(\boldsymbol{\tau}_j) \\
m_j^* &= \frac{\alpha m_j + n\bar{\xi}_j}{\alpha + n_j}, \\
\alpha_j^* &= \alpha_j + n_j, \\
a_j^* &= a + n_j, \\
b_j^* &= b + (n_j - 1)s_j^2 + \frac{\alpha n_j}{\alpha + n_j}(m_j - \bar{\xi}_j)^2,
\end{aligned}$$

RWMH

Likelihood

$$L_{c-vine} = \sum_{trees} \sum_{edges} L(edge)$$

Induced priors

$$f_\tau(\tau) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{ -\frac{1}{2} \cdot \left(\frac{\log\left(\frac{\tau}{1-\tau}\right) - \mu}{\sigma} \right)^2 \right\} \cdot \frac{1}{\tau \cdot (1-\tau)}, \quad \tau \in (0,1)$$

$$f_\tau(\tau) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{ -\frac{1}{2} \cdot \left(\frac{\log\left(\frac{1+\tau}{1-\tau}\right) - \mu}{\sigma} \right)^2 \right\} \cdot \frac{2}{1-\tau^2}, \quad \tau \in (-1,1).$$

ABC

Algorithm proposed by Beaumont et al. (2002).

- Simulate M 10-dimensional vectors $\phi_j \sim \mathcal{G}\left(\frac{a_j}{2}, \frac{b_j}{2}\right)$ and $\mu_j \sim \mathcal{N}\left(m_j, \frac{1}{\alpha_j \cdot \phi_j}\right)$.

$$\left((\mu_{1(1)}, \phi_{1(1)}), \ldots, (\mu_{10(1)}, \phi_{10(1)})\right)$$

$$\vdots$$

$$\left((\mu_{1(M)}, \phi_{1(M)}), \ldots, (\mu_{10(M)}, \phi_{10(M)})\right)$$

- Simulate 12 values $\xi_j^k \sim N(\mu_j, \frac{1}{\phi_j})$, for each pair (μ_j, ϕ_j).

$$(\xi_{1(1)}^{1}, \ldots, \xi_{1(1)}^{12}, \xi_{2(1)}^{1}, \ldots, \xi_{10(1)}^{12})$$

$$\vdots$$

$$(\xi_{1(M)}^{1}, \ldots, \xi_{1(M)}^{12}, \xi_{2(M)}^{1}, \ldots, \xi_{10(M)}^{12})$$

- Transform $\boldsymbol{\xi}$ into $\boldsymbol{\tau}$ and $\boldsymbol{\tau}$ into $\boldsymbol{\theta}$.

$$(\theta_{TH(1)}^{00.1}, \ldots, \theta_{TH(1)}^{11.3}, \theta_{TR(1)}^{00.1}, \ldots, \theta_{PD|THR(1)}^{11.3})$$

$$\vdots$$

$$(\theta_{TH(M)}^{00.1}, \ldots, \theta_{TH(M)}^{11.3}, \theta_{TR(M)}^{00.1}, \ldots, \theta_{PD|THR(M)}^{11.3})$$

ABC

- Simulate values from each one of these c-vines, with these parameters.

$$\mathbf{u}_{(1)} = (u_{1(1)}, \ldots, u_{5(1)})$$
$$\vdots$$
$$\mathbf{u}_{(M)} = (u_{1(M)}, \ldots, u_{5(M)})$$

- Obtain the Kendall$-\tau$ between each pair of c-vine nodes in each simulation, that is, the proposed parameters $\rightarrow (\boldsymbol{\tau_{(i)}}, S_{(i)}), i = 1, \ldots, M$.

$$\mathbf{S}_{(1)} = (S^{00.1}_{TH(1)}, \ldots, S^{11.3}_{PD|THR(1)})$$
$$\vdots$$
$$\mathbf{S}_{(M)} = (S^{00.1}_{TH(M)}, \ldots, S^{11.3}_{PD|THR(M)})$$

- Calculate the Euclidean distance between S and each $S_{(i)}$, $\|S_i - S\| = \sqrt{\sum_{j,k}(s^k_{j(i)} - s^k_j)^2}$, and select the subset of $(\boldsymbol{\tau_{(i)}}, S_{(i)})$ whose Euclidean distance to S is smaller than a tolerance threshold.
- Denote as i^* the index of the selected parameters.

ABC

- Perform a local linear regression between $S_{(i^*)}$ and S.
- The sample from the posterior parameter distribution: $\hat{\tau}_{(i^*)} = \tau_{(i^*)} - (S_{(i)} - S)^T \hat{\beta}$,
- $(\mu_{(i^*)}, \phi_{(i^*)})$ is a sample of the posterior distribution of the hyperparameters.

ABC vs. RWMH

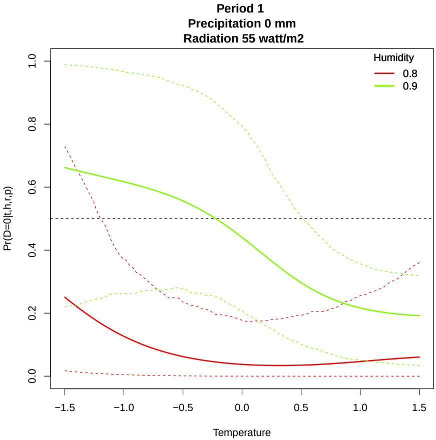
τ		RWMH	ABC
	True value	mean (95% cred int.)	mean (95% cred int.)
TH	0.317	0.324 (0.297,0.353)	0.279 (0.232,0.322)
TR	0.561	0.565 (0.524,0.606)	0.528 (0.504,0.550)
TP	0.776	0.783 (0.771,0.795)	0.751 (0.706,0.794)
$HR \mid T$	-0.193	-0.235 (-0.299,-0.172)	-0.209 (-0.247,-0.173)
$HP \mid T$	0.675	0.670 (0.643,0.695)	0.632 (0.596,0.667)
$RP \mid TH$	0.622	0.621 (0.579,0.661)	0.638 (0.600,0.677)

Execution times of RWMH and ABC algorithms

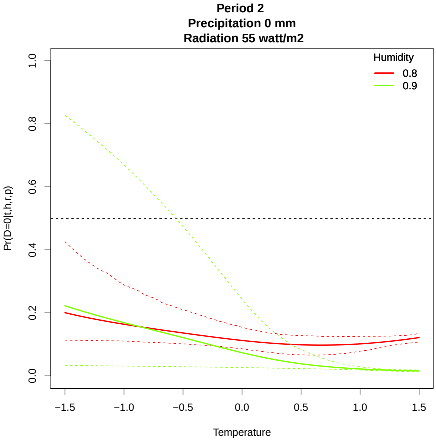
		Time	
Length	Acep./tol.	RWMH	ABC
10,000	20%	$1h.12'26''$	$44'31''$
30,000	20%	$3h.24'33''$	$2h.08'38''$
50,000	20%	$5h.24'27''$	$3h.44'31''$

Obtained with a desktop computer with a Intel(R) Core(TM) i5-330M CPU@2.60GHz processor

Results

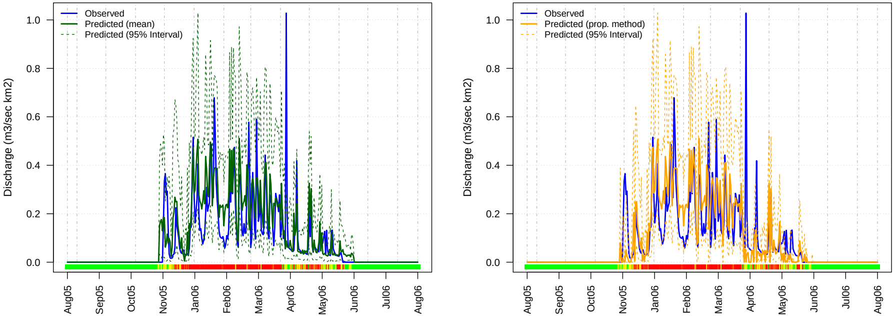


Period 1
Precipitation 0 mm
Radiation 55 watt/m2
Humidity
0.8
0.9
Pr(D=0|t,h,r,p)
0.0
0.2
0.4
0.6
0.8
1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Temperature



Period 2
Precipitation 0 mm
Radiation 55 watt/m2
Humidity
0.8
0.9
Pr(D=0|t,h,r,p)
0.0
0.2
0.4
0.6
0.8
1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Temperature

Results


Observed
Predicted (mean)
Predicted (95% Interval)
Discharge (m3/sec km2)
0.0
0.2
0.4
0.6
0.8
1.0
Aug05
Sep05
Oct05
Nov05
Jan06
Feb06
Mar06
Apr06
May06
Jun06
Jul06
Aug06
Observed
Predicted (prop. method)
Predicted (95% Interval)
Discharge (m3/sec km2)
0.0
0.2
0.4
0.6
0.8
1.0
Aug05
Sep05
Oct05
Nov05
Jan06
Feb06
Mar06
Apr06
May06
Jun06
Jul06
Aug06

Summary

- Multivariate copula model
  - Temperature, humidity, radiation, precipitation and discharge
  - c-vine copula model
  - Seasonality though division in periods
- Hierarchical model
  - Hierarchical structure over groups and periods
  - ABC algorithm to make inference
  - Comparison between ABC and RWMH

Current and future research

Next steps in our research:

- Construct a hierarchical model for the marginal and copula parameters
- Analyze the influence of order of the variables in the c-vine
- Explore other vine structures

Further extensions of this work:

- Update models with data registered but not still collected by GLACKMA
- Develop similar models for glaciers sited in other latitudes and altitudes
- Adding of new meteorological variables as the direction and speed of wind

Main references

- Gómez, M., Ausín, M. C., and Domínguez, M. C. (2016). Vine copula models for predicting water flow discharge at King George Island, Antarctic. *UC3M Working Papers, Statistics and Econometrics*. 16-12.
- Domínguez M.C. (2004). Software for gauging. VI Symposium Glacier Caves and Karst in Polar Regions. Ny-Alesund (Svalbard), Norway, 27-36.
- Aas, K., Czado, C., Frigessi, A., Bakken, H., (2009). Pair-copula constructions of multiple dependence, *Insurance: Mathematics and Economics* 44 (2), 182–198.
- Hock, R. (2005). Glacier melt: A review of processes and their modelling. *Progress in Physical Geography*, 29, 362–391
- Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). *Approximate Bayesian computation in population genetics*. Genetics, 162(4):2025-2035.
- Bedford, T. and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial intelligence*, 32(1–4):245-268.

Decomposition with the c-vine structure

$$f(h \mid t) = \frac{f(t,h)}{f_t(t)} = \frac{c_{th}(F_t(t), F_h(h)) \cdot f_t(t) \cdot f_h(h)}{f_t(t)}$$
$$= c_{th}(F_t(t), F_h(h)) \cdot f_h(h)$$

$$f(r \mid t,h) = \frac{f(h,r \mid t)}{f(h \mid t)}$$
$$= \frac{c_{hr|t}(F(h \mid t), F(r \mid t)) \cdot f(h \mid t) \cdot f(r \mid t)}{f(h \mid t)}$$
$$= c_{hr|t}(F(h \mid t), F(r \mid t)) \cdot f(r \mid t)$$
$$= c_{hr|t}(F(h \mid t), F(r \mid t)) \cdot c_{tr}(F_t(t), F_r(r)) \cdot f_r(r)$$

Decomposition with the c-vine structure

$$
\begin{aligned}
f(p \mid t,h,r) &= \frac{f(r,p \mid t)}{f(r \mid t,h)} \\
&= \frac{c_{rp|th}\left(F(r \mid t,h), F(p \mid t,h)\right) \cdot f(r \mid t,h) \cdot f(p \mid t,h)}{f(r \mid t,h)} \\
&= c_{rp|th}\left(F(r \mid t,h), F(p \mid t,h)\right) \cdot f(p \mid t,h) \\
&= c_{rp|th}\left(F(r \mid t,h), F(p \mid t,h)\right) \cdot c_{hp|t}\left(F(h \mid t), F(p \mid t)\right) \\
&\cdot c_{tp}\left(F_t(t), F_p(p)\right) \cdot f_p(p)
\end{aligned}
$$

$$
\begin{aligned}
f(d \mid t,h,r,p) &= \frac{f(p,d \mid t)}{f(p \mid t,h,r)} \\
&= \frac{c_{pd|thr}\left(F(p \mid t,h,r), F(d \mid t,h,r)\right) \cdot f(p \mid t,h,r) \cdot f(d \mid t,h,r)}{f(p \mid t,h,r)} \\
&= c_{pd|thr}\left(F(p \mid t,h,r), F(d \mid t,h,r)\right) \cdot f(d \mid t,h,r) \\
&= c_{pd|thr}\left(F(p \mid t,h,r), F(d \mid t,h,r)\right) \cdot c_{rd|th}\left(F(r \mid th), F(d \mid t,h)\right) \\
&\cdot c_{hd|t}\left(F(h \mid t), F(d \mid t)\right) \cdot c_{td}\left(F_t(t), F_d(d)\right) \cdot f_d(d)
\end{aligned}
$$