Stochastic Downscaling of Precipitation with Conditional Mixture Models

Julie Carreau
HydroSciences Montpellier, julie.carreau@univ-montp2.fr

Mathieu Vrac
Laboratoire des Sciences du Climat et de l'Environnement

1 Introduction

Statistical downscaling models (SDMs) seek to bridge the gap between large-scale variables simulated from General Circulation Models (GCMs) and small scale variables with high spatial variability such as precipitation. In this paper, we propose to model the distribution of precipitation conditional on large-scale atmospheric information with conditional mixture models (CMMs). CMMs are mixture models whose parameters are computed by a neural network based on large-scale atmospheric predictors. We consider three types of CMMs which differ in the type of continuous densities (Gaussian, Log-Normal or hybrid Pareto) they use as mixture components. We evaluate the three CMMs against the two-component mixture from Williams [3] at downscaling precipitation at three rain gauge stations in the French mediterranean area.

2 Materials and Methods

CMMs combine a discrete component for the "no rain" events and a continuous component for rainfall intensity and can be written as :

$$\phi(y;\psi) = \underbrace{(1-\alpha)\delta(y)}_{\text{no rain}} + \underbrace{\alpha\phi_0(y;\psi_0)}_{\text{rain>0}}, \tag{1}$$

where α is the rain probability, $\delta(\cdot)$ is the Dirac function, $\phi_0(\cdot;\psi_0)$ is the density for rainfall intensity with parameter ψ_0 and $\psi = (\alpha, \psi_0)$. In [3], $\phi_0(\cdot;\psi_0)$ is the Gamma density. We propose to use mixtures instead. We can take into account the dependence of the distribution of precipitation on large-scale atmospheric variables by considering the parameters of the mixture as functions of these variables. A convenient way to implement these functions is by means of a neural network (NN) [1]. The NN parameters are calibrated by minimizing the negative log-likelihood of the conditional mixture over the training set. We selected the hyper-parameters (the number of hidden units and the number of components) via the cross-validation method, see [1]. We evaluate three CMMs which differ in the type of mixture components and compare them with the two-component mixture from Williams [3]. We took Gaussian, Log-Normal or hybrid Pareto ([2]) as mixture components.

The local-scale data are precipitation from three rain gauge stations, Orange, Sète and Le Massegros which are located in the Cévennes-Vivarais, in the French Mediterranean area. Because of the Mediterranean influence and of the mountainous back country, the Cévennes-Vivarais region is well known for intense rain events, especially in the fall. We have daily rainfall measurements over 46 years (01/01/1959 -12/31/2004) from the *European Climate Assessment & Dataset* (ECA&D). The set of predictors includes the NCEP/NCAR (National Centers for Environmental Prediction/National Center for Atmospheric Research) reanalysis sea level pressure (SLP) fields on a 6 by 6 grid cell regions surrounding the stations. We also include as predictors three date variables representing the year, the month and the week of an observation. Principal component analysis is applied to reduce the dimensionality and remove the redundancy among the predictors. We extract the four principal components in order to keep 90% of the variance of the data.

The 46-year data set is split into a training set of 25 years (01/01/59 - 12/31/83) and a test set of 21 years (01/01/84 - 12/31/04). The training set is first used to select the hyper-parameters with the 5-fold cross-validation method. Then, each model is trained anew on the whole training set with the selected hyper-parameters. The test set serves exclusively for comparison and evaluation of the SDMs.

3 Results

The hybrid Pareto CMM being the most complex model, we first compare the other three SDMs in terms of relative log-likelihood with the hybrid Pareto CMM on the test set. Table 1 shows the relative log-likelihood on the test set along with standard errors for the three competing SDMs on the three rain gauge stations. In bold font are the cases where the hybrid Pareto CMM performed significantly better. We see that the hybrid Pareto CMM outperforms the Gaussian CMM and the Gamma benchmark on all three stations. However, we cannot really distinguish the hybrid Pareto CMM from the Log-Normal CMM based on this criterion.

	Gaussian	Log-Normal	Williams
Orange	**0.02146 (0.003139)**	0.0022512 (0.001910)	**0.02275 (0.002866)**
Sète	**0.01595 (0.003034)**	-0.003530 (0.001647)	**0.01847 (0.002690)**
Le Massegros	**0.01948 (0.006671)**	-0.004606 (0.002121)	**0.02068 (0.003005)**

Table 1: *Relative log-likelihood (std. err.) on the test set between the hybrid Pareto CMM and the other SDMs (Gaussian and Log-Normal CMMs and Williams' model). Positive numbers indicate that the hybrid Pareto CMM performed better. Significant differences are in bold font.*

We randomly generated data for each SDM corresponding to the predictor values on the test set. This was repeated a thousand times. Fig. 1 illustrates the QQ-plots for Orange, on logarithmic scale, between the observations and the simulations for the hybrid Pareto CMM, left panel, and for Williams' model (right panel). Models which are in accordance with the data should be close to the diagonal line. We see that Williams' model is less apt at modelling both the central part (over-estimation) and the upper part (under-estimation) of the distribution. In Fig. 2, we first analyze the seasonal cycles of the rain probability (left panel) and of the 99% quantile (right panel) of the hybrid Pareto CMM on the Orange test set. We can identify from Fig. 2 two seasonal modes, around March (03) and October (10), which translates into higher probabilities and amounts of rain around these two months, while summer (i.e., around July) presents lower probabilities and amounts of rain. This is globally in agreement with the observations over the test set, showing the same features. In Fig. 3, we finally look at the conditional densities of the hybrid Pareto CMM associated with different atmospheric conditions, that is for different predictors, for the rain event at the Orange station with the highest volume of rain (322 mm in 09/08/2002-09/09/2002) in the test set. The left panel of Fig. 3 shows the central part of the conditional densities while the right panel represents the upper tails in logarithmic scale. Each curve corresponds to a different day which is connected in the legend with the amount of rain observed on that day in chronological order (from top to bottom). From Fig. 3, we see that the conditional density is very responsive to changes in atmospheric conditions and that globally, days with heavy rains correspond to heavy tailed densities and days with no rain to almost flat densities.

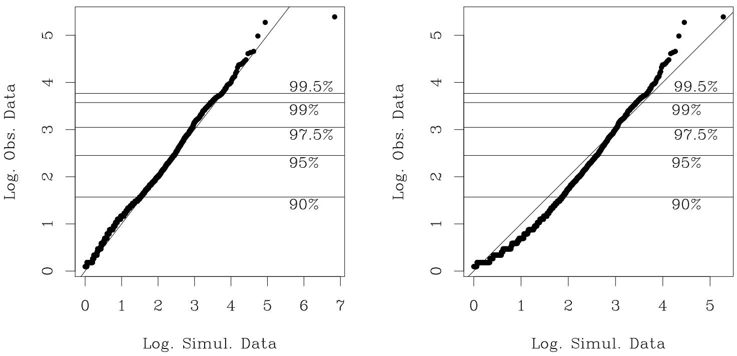

Figure 1: *QQ-plots on logarithmic scale of the simulated precipitation versus observations $> 1mm$ on the Orange test set for the hybrid Pareto CMM (left panel), and Williams model (right panel). The horizontal lines are the empirical unconditional quantiles from observations of the test set.*

4 Concluding remarks

To our knowledge, CMMs are used for the first time in a downscaling context and open interesting ways to study the interactions between large- and small-scale climate variables. CMMs extend the two-component mixture proposed initially by Williams [3] which has a discrete component like CMMs to model rainfall occurrence but relies on a single density, the Gamma, for rainfall intensity.

We draw the following conclusions from our analyses on the three stations in the French mediterranean area: 1) CMMs have clear advantages over Williams model in terms of flexibility to represent both the central and the extremal part of rainfall intensity distribution and 2) the choice of component in CMMs depends on the data. In our case, Gaussian

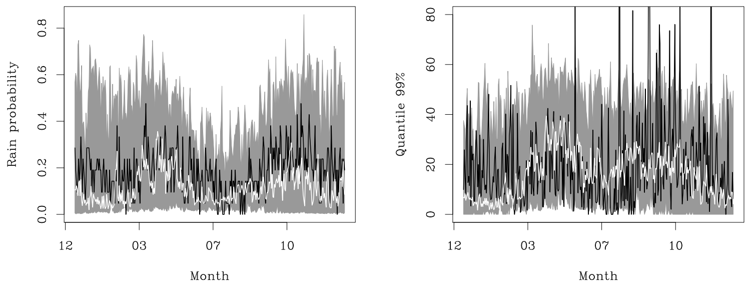

Figure 2: *Daily seasonal cycles of the rain occurrence probability (left panel) and of the 99% quantile (right panel) from the observations (black line) together with an empirical 90% confidence interval (grey band) and median (white line) from the hybrid Pareto CMM for the Orange station test data.*

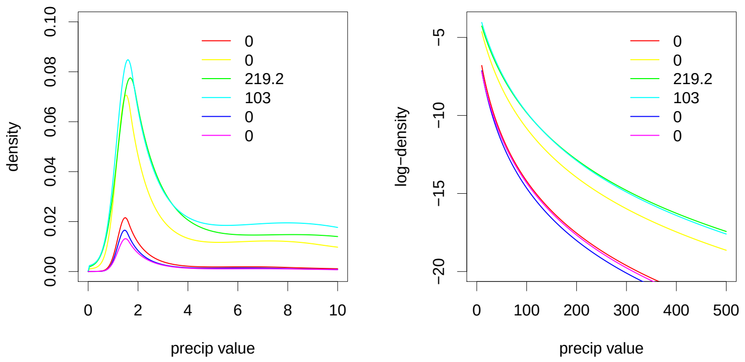

Figure 3: *Conditional densities for the hybrid Pareto CMM day by day for a period with the highest volume of rain in the test Orange data. Each daily density is represented with a different color which is represented in the legend in chronological order, from top to bottom, with the amount of rainfall observed.*

components are not well suited. Log-Normal CMMs offer a good performance and are more straightforward to implement than hybrid Pareto CMMs. However, the assumption of heavy tails of the hybrid Pareto CMM seems more realistic for the precipitation data considered in this work.

References

[1] C. Bishop. *Neural Networks for Pattern Recognition.* Oxford, 1995.

[2] J. Carreau and Y. Bengio. A hybrid Pareto model for asymmetric fat-tailed data: the univariate case. *Extremes*, 12(1):53–76, 2009.

[3] M. P. Williams. Modelling seasonality and trends in daily rainfall data. In *Advances in Neural Information and Processing Systems*, volume 10, pages 985–991, 1998.