Spatio-Temporal Censored Model of Precipitation Climatology

Reto Stauffer[1], Nikolaus Umlauf[1], Jakob W. Messner[1], Georg J. Mayr[2], Achim Zeileis[1]

[1] Department of Statistics, University of Innsbruck, Austria
[2] Institute of Meteorology and Geophysics, University of Innsbruck, Austria

E-mail for correspondence: `reto.stauffer@uibk.ac.at`

Abstract: Flexible spatial-statistical models are widely used to create climatological estimates. Although most models assume a normally distributed response this assumption can lead to inaccurate estimates for certain variables. Precipitation, for instance, is physically limited to values ≥ 0 so that it might be seen as left-censored. We develop a novel spatial-statistical additive model for location, scale, and shape which can handle censored normal distributed responses. This article presents a precipitation climatology over complex terrain with a daily temporal resolution on a 0.5 x 0.5 km grid. The results demonstrate that the new method outperforms existing methods and is able to resolve local effects quite well compared to single-station estimates. Our model enables the creation of climatologies with a high spatio-temporal resolution, yet does not need extensive tuning. Overall, the model is easily adaptable to new applications.

Keywords: spatial climatology; spatial modelling; precipitation; censoring.

1 Introduction

The northern part of Tyrol in the Eastern Alps is reaching from $465m$ up to $3798m$, the highest peak in Austria. Due to the complex topography and the location within the Alps, Tyrol is strongly related to precipitation. During the winter the local economy, tourists, transportation, and the local population is strongly related to snow conditions while during the summer season extreme precipitation amounts can lead to floods or droughts.

Using spatial-statistical models to create climatologies is not new. However, existing methods are often based on monthly or even yearly means or sums. Plausible reason: there is no need to handle zero-observation events and the

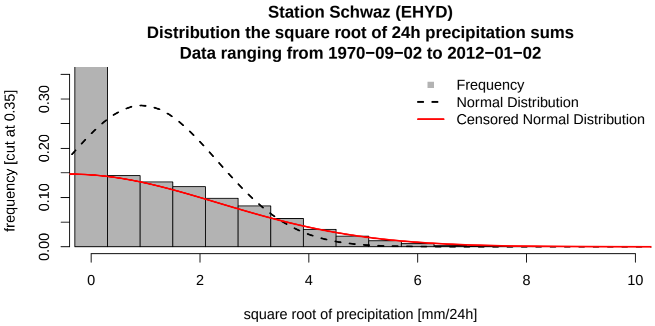

FIGURE 1. Sample distribution of 24h precipitation sums at station Schwaz (535 $m.a.s.l$, Tyrol; observations square-root transformed) containing 14975 observations. The lines show the PDFs for a normal distribution (black, dashed), and for a normal distribution left-censored at 0 (red, solid). Ordinate cut at 0.35. Class $[-0.3, 0.3]$ has a frequency of 0.95.

assumption of a normally distributed responses is often apposite. On a daily base zero-observations occur very frequently. We developed an extended additive model for location, scale, and shape for censored normally distributed response. The field of possible applications is reaching from climatological trend analysis, torrent control, or local spatial planning, to the use as priori information for extreme-value modelling or weather forecasting.

2 Skewness of Observation Distribution

Precipitation observations are typically skewed distributed. A common approach to remove major parts of the skewness is to transform the observed values as proposed by Cox (1964). For precipitation, a square root transformation is commonly used (Hutchinson 1998b). Furthermore the data show a strong point mass on 0 due to the fact that about two third of all days are “dry” without precipitation. As precipitation can only be ≥ 0, the distribution can be seen as left-censored at 0 (Cohen, 1959).
Figure 1 shows the distribution of observed square root transformed daily precipitation sums for station Schwaz. Additionally, two probability distribution functions (PDFs) are shown. The first one shows a normal distribution with in-sample mean and standard deviation, the second shows a left-censored normal distribution with its censoring point at 0. The latter one is obviously more appropriate for the given sample distribution.

3 Censored Spatial Model

For the spatial modelling we are using a new R package called `bamlss` (Umlauf 2015), a package for Bayesian Additive Models for Location, Shape and Scale. The model can be written as:

$$y^* \sim N(\mu, \sigma^2)$$

Where y^* is the latent response, η is the location, and σ the scale parameter of the latent unobservable process. Location and log-scale are expressed by an additive combination of explanatory variables. In the "simple" example shown in Figure 2 & 3 these are: `yday` (day of the year), `alt` (altitude), and `lon/lat` (longitude/latitude).

$$\eta_\bullet = \varphi_0 + g(\text{yday}) + g(\text{alt}) + g(\text{lon}, \text{lat})$$

where $\varphi_0 = \beta_0$ for η_{mu}, $\varphi_0 = \gamma_0$ for η_σ, and g are non-linear functions determined by coefficients β and γ respectively.

$$\mu = \eta_\mu; \quad \log(\sigma) = \eta_\sigma; \quad \mathbf{y} = \max(\mathbf{0}, \mathbf{y}^*)$$

Given the response y, and the explanatory variables x, the unknown parameters for η_μ (β) and for η_σ (γ) can be found by maximum likelihood:

$$L(\beta, \gamma | y, x) = \prod_{i=1}^{N} f(y_i | x_i, \beta, \gamma)^{I(y_i > 0)} \cdot F(0 | x_i, \beta, \gamma)^{I(y_i = 0)}$$

where f and F are the probability density function and the cumulative distribution function of the normal distribution respectively.
Preliminary results can be found in Figure 2 & 3 showing the effects included in the model formulas $\eta_\bullet$. The 110 stations used for the model are reaching up to $2290m$ but, about 90 % are located below $1600m$ which leads to an increasing log-scale with higher altitude as shown in Figure 2.

4 Model Comparison & Intermediate Findings

Figure 4 shows a comparison of the seasonal effects between single-station estimates (colorized) and the mean seasonal effect on the full spatial model (black). The seasonal effect strongly varies between the south and the north side of the Alps. The spatio-temporal model captures the characteristic of the major proportions, however, the south side stations (reddish colors) should have a different effect. To take this into account an additional effect (e.g. spatially variable seasonal effects) should be added but has not been done yet.
To check the improvement of the new censored spatio-temporal estimates, we compared the results of the censored models (using bamlss) against some

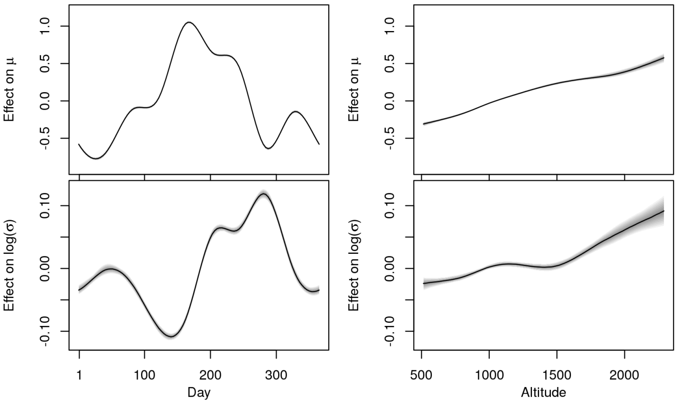

FIGURE 2. Centred effects for location (top row) and log-scale (bottom row) on square root transformed observations. Left/Right: effect on day of the year/altitude. Effects estimated on 30 years of data (1982-2012), 110 different stations. Highest precipitation amounts expected during summer period (μ). Log-scale seems to be highest in late summer/autumn (day 200–300; late July to September) which is related to the convective season.

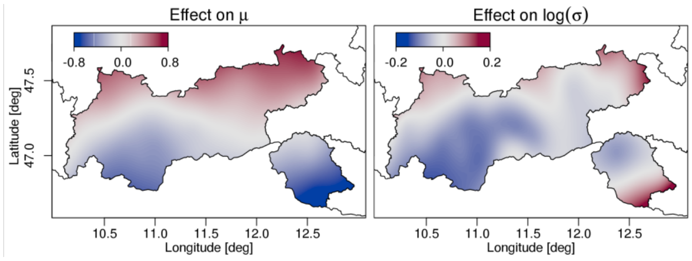

FIGURE 3. Centred spatial longitude/latitude effect over North Tyrol for location (left), and log-scale (right). Same data basis as Figure 2. The location effect shows the increasing amounts of precipitation towards the northern parts (Alpine ridge). Lowest in dry inner Alpine valleys. The log-scale effect seems to be strongly related to the station density (boundary-effect).

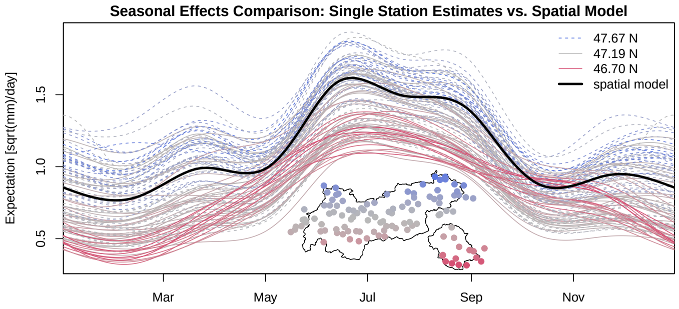

FIGURE 4. Each line plotted shows the climatology for one of the stations estimated via a single-station based censored bamlss model. The colors correspond to the latitudinal location of the station (see subplot; overview over North Tyrol). In black: mean effect from the spatio-temporal model shown in Figure 2 & 3.

generalized additive models based on monthly mean observations (using mgcv). As the temporal resolution of these models differ (daily vs. monthly) two questions arise: (i) What is the skill of the new daily-based climatology when aggregated to monthly means vs. a model directly trained on monthly means? (ii) How is the skill on a daily basis?

The in-sample validation of four selected models is shown in Figure 5. Skill scores of mean absolute and root mean squared errors are shown. Model “mon-A/cens-A” and “mon-B/cens-B” are comparable as they have identical explanatory effects. The setup of model “A” is shown in Chapter 3, setup “B” contains additional topographic regressors.

As shown our new model is comparable to the one estimated on monthly mean values when aggregated to a monthly resolution (Figure 5). On a daily basis the new model has significantly lower errors (positive skill).

5 Conclusion & Outlook

The presented method has some major advantages over existing ones. Treating precipitation as left censored makes it unnecessary to reduce the temporal resolution (e.g., take monthly means to avoid zero-observations). Furthermore additional daily-based explanatory variables can be included to create e.g, wind-direction dependent climatological estimates. In comparison to commonly used tools for precipitation analysis, the new approach needs less priori knowledge and less extensive tuning. Furthermore, the new method outperforms existing ones even when aggregated to a monthly resolution.

As shown, not all effects are captured yet. Searching for additional variables has to be one of the next steps. Due to the size of the data set additional (multi-dimensional) terms are computationally expensive (memory

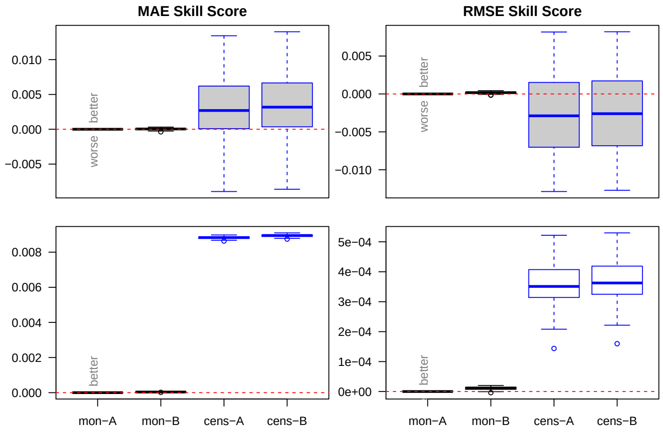

FIGURE 5. Skill scores of bootstrapped mean absolute (MAE) and root mean squared (RMSE) errors validated on monthly-mean base (top row), and on a daily basis (bottom row). Model “mon-A” chosen as reference model. Black: normal distributed models on monthly means (mgcv). Blue: censored normal distributed models on daily observations (bamlss). The unit of all data is in “square root of daily mean” ($\sqrt{millimeter}$ per day).

& CPU). Therefore we are working on the code of the bamlss package to optimize the performance/resource management.

Acknowledgments: The ongoing project is supported by the Austrian Science Fund (FWF): TRP-290.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 211 – 252.

Cohen, A. C. J. (1959). Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*, **1 (3)**, 99 – 113.

Hutchinson, M. F. (1998a). Interpolation of rainfall data with thin plate smoothing splines – Part I. *Journal of Geographic Information and Decision Analysis*, **2**, 168 – 185.

Hutchinson, M. F. (1998b). Interpolation of rainfall data with thin plate smoothing splines – Part II. *Journal of Geographic Information and Decision Analysis*, **2 (2)**, 152 – 167.

Umlauf, N. (2015). A conceptional Lego toolbox for Bayesian distributional regression models. *30th International Workshop on Statistical Modelling.*.