Nat. Hazards Earth Syst. Sci., 15, 1757–1761, 2015
www.nat-hazards-earth-syst-sci.net/15/1757/2015/
doi:10.5194/nhess-15-1757-2015

Natural Hazards and Earth System Sciences
Open Access

Statistical detection and modeling of the over-dispersion of winter storm occurrence

M. Raschke

Independent researcher, Stolze-Schrey-Str. 1, Wiesbaden, Germany

Correspondence to: M. Raschke (mathiasraschke@t-online.de)

Received: 1 February 2015 – Published in Nat. Hazards Earth Syst. Sci. Discuss.: 3 March 2015
Revised: 28 July 2015 – Accepted: 29 July 2015 – Published: 11 August 2015

Abstract. In this communication, I improve the detection and modeling of the over-dispersion of winter storm occurrence. For this purpose, the generalized Poisson distribution and the Bayesian information criterion are introduced; the latter is used for statistical model selection. Moreover, I replace the frequently used dispersion statistics by an over-dispersion parameter which does not depend on the considered return period of storm events. These models and methods are applied in order to properly detect the over-dispersion in winter storm data for Germany, carrying out a joint estimation of the distribution models for different samples.

1 Introduction

A possible over-dispersion of the occurrence of European winter storms is the topic of previous studies (e.g., Mailier et al., 2006; Pinto et al., 2013; Karremann et al., 2014). Over-dispersion means that the variance (a spreading parameter) of the number of events per storm season is larger than the corresponding expectation. This is frequently called clustering. However it is difficult to detect statistical significance of over-dispersion of historic winter storms due to the small sample sizes and the relatively small over-dispersion (cf. Vitolo et al., 2009).

Here I present a combination of statistical models and methods that improve the modeling of the over-dispersion of European winter storms and the detection of its statistical significance. The generalized Poisson distribution (GPD) is introduced in the following section. It is more universal than the negative binomial distribution (NBD) which is already applied to storms (e.g., Karremann et al., 2014). In Sect. 3, I introduce an over-dispersion parameter which remains the same for each return period, in contrast to frequently used dispersion statistics (e.g., Karremann et al., 2014). The corresponding thinning process is also explained. In Sect. 4, I introduce the statistical model selection by an information criterion. Finally, I use this criterion when I apply the GPD to the data set of German winter storms analyzed in Karremann et al. (2014). Therein the over-dispersion in the data of historic storms is well detected by an over-dispersion parameter, estimated for a large sample of storms from climate model simulations.

2 The generalized Poisson distribution (GPD)

The occurrence of European winter storms on a timescale can be modeled by an inhomogeneous Poisson process (Mailier et al., 2006). In consequence, the number of storms per season, with a magnitude equal to or larger than a defined level, can be modeled by a distribution that is a mixture of Poisson distributions (PDs). The NBD is such a mixture and is already applied to storm occurrence (e.g., Sakamoto, 1973; Karremann et al. 2014). The PD is a limit case of the NBD and includes neither over- nor under-dispersion. In contrast to the limitations of the NBD to the case of over-dispersion (cf. Johnson et al., 2005, Eq. 5.5), the GPD can consider over-dispersion or under-dispersion and includes the PD as a special case. The GPD is more universal and it also is a mixture of PD in case of over-dispersion (Joe and Zhu, 2005). Hence I use the GPD. The GPD has been developed by Consul and Jain (1973) and is formulated for the discrete random variable $X \geq 0$ for count data with

$$P(x) = \frac{\lambda(\lambda + x\theta)^{x-1}}{x!} e^{-\lambda - x\theta}, \quad x \geq 0. \tag{1}$$

Published by Copernicus Publications on behalf of the European Geosciences Union.

It describes the probability of $X = x$ and uses the parameters $\lambda > 0$ and θ. The expectation $E(X)$ and variance $V(X)$ can replace the parameters λ and θ having

$$\theta = 1 - \sqrt{\frac{E(X)}{V(X)}} \tag{2}$$

and

$$\lambda = \sqrt{\frac{E(X)^3}{V(X)}}. \tag{3}$$

A certain inhomogeneous Poisson process of event occurrence in time does not necessarily need to result exactly in a NBD or a GPD for the count data of events per storm season. But I assume that the GPD is at least an appropriate approximation. The similarity of the NBD and the GPD for the case of over-dispersion supports this assumption; each approximates the other. This is later validated for and by the analyzed storm data.

The parameters of the GPD can be estimated by the well-known maximum likelihood (ML) method. Therein the parameter vector $\boldsymbol{\theta}$, with the maximum of the logarithmized likelihood function

$$\log(L(\boldsymbol{\theta})) = \sum_{x}^{n_x} \log(P(x;\boldsymbol{\theta})) \tag{4}$$

represents the point estimation. n_x is the observed number of seasons with $X = x$. The ML method is recommended for the NBD (Johnson et al., 2005) and the GPD (Consul and Shoukri, 1984). Therein, the estimated expectation $\hat{E}(X)$ is equal to the sample mean. The moment method is also an established estimation method for discrete distributions in contrast to the least square method (cf. Johnson et al., 2005).

3 What does over-dispersion mean?

Mailier et al. (2006), Vitolo et al. (2009) and Karremann et al. (2014) define the over-dispersion in the random distribution of the number of storms as serial clustering, and quantify it by the dispersion statistics

$$\phi = V(X)/E(X) - 1. \tag{5}$$

However, the terms cluster and clustering have different meanings. A cluster in an auto-correlated time series consists of a number of observations that represent a partial series of exceedances of a given threshold (Coles, 2001, Fig. 5.4), e.g., of the time series of river discharge. An earthquake cluster is a group of earthquake events that include a main event with a large magnitude and a series of secondary events (after- and/or foreshocks; e.g., Ogata, 2001). There is a relation in time and space between these events. If a clustering of storm magnitudes existed, similar to the clustering of earthquakes, the number of smaller storms with low return periods (RPs) should be higher for years with an event with a high RP. Nevertheless, this cannot be stated for the data analyzed by Karremann et al. (2014). This fact is associated with the physical processes leading to the occurrence of cyclone clusters (Pinto et al., 2014).

Here, X is the number of storms, in a season, whose storm magnitudes have RPs equal to or larger than a considered return period (CRP). If over-dispersion in the count variable X were generated only by an inhomogeneous Poisson process at the timescale, then each storm event would be independent of the others. In this case, the increase of the CRP from $\mathrm{CRP}_{\mathrm{old}}$ to $\mathrm{CRP}_{\mathrm{new}}$ includes a random thinning of the sampling of X. Therein, the survival probability $P_{\mathrm{survival}} = \mathrm{CRP}_{\mathrm{old}}/\mathrm{CRP}_{\mathrm{new}}$ with $\mathrm{CRP}_{\mathrm{old}} > \mathrm{CRP}_{\mathrm{new}}$ is the probability of a storm event of the sample with $\mathrm{CRP}_{\mathrm{old}}$ to be also member of the sample with a higher $\mathrm{CRP}_{\mathrm{new}}$. The RP of the magnitude of a storm, which is also a random variable, determines if this storm stays in the sample with $\mathrm{CRP}_{\mathrm{new}}$ or not. In other words, the counted storm events are thinned out by the increasing of the CRP. A possible over-dispersion with $V(X) > E(X)$ is determined independent of the CRP by (cf. Mack, 2002)

$$V(X) = E(X) + \beta E(X)^2, \tag{6}$$

where the over-dispersion parameter $\beta \geq 0$ ($\beta = 0$ means no over- or under-dispersion). This is a result of well-known relations between random variables (cf. Ross, 2007). I present a derivation of Eq. (6) in the Supplement. The relation between the expectation of the storm number and the CRP is obviously $E(X) = 1/\mathrm{CRP}$. The important advantage of the over-dispersion parameter β is that it has one fixed value for all CRP, in contrast to the over-dispersion statistics ϕ which depends on the CRP (cf. Karremann et al., 2014; they use the term *return level*). An example of the corresponding relations is shown in Fig. 1. The over-dispersion statistics ϕ decreases with the increasing of $\mathrm{CRP} = 1/E(X)$ for the fixed β. This behavior explains very well, as by Karremann et al. (2014; Table 3), the estimations for large storm samples from climate model simulations. It seems to be likely that the over-dispersion of storms is caused by inhomogeneous occurrence intensity of events in time.

4 Statistical significance and model selection

Statistical significance means that a certain assumption and/or modeling is likely appropriate and correct. It is frequently insured by a test, such as the likelihood ratio test, in the model selection. The statistical significance of a model can also be provided indirectly by an appropriate statistical model selection among different alternatives by means of an information criterion. It is recommended, for example, for regression analysis with binary variables instead of a clas-

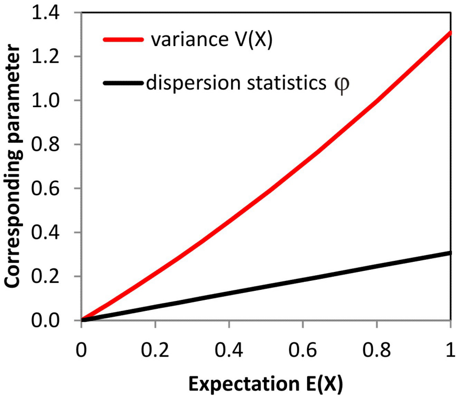

Figure 1. Relation between variance and expectation according to Eq. (6) with dispersion parameter $\beta = 0.307$ and corresponding behavior of dispersion statistics ϕ according to Eq. (5) ($\mathrm{CRP} = 1/E(X)$).

sical t test (Fahrmeir et al., 2013). The functionality of such model selection can also be validated by Monte Carlo simulations (Raschke and Thürmer, 2010). Here, I apply for model selection the well-known Bayesian information criterion of Schwarz (1978):

$$\mathrm{BIC} = -2L(\boldsymbol{\theta}) + m\log(n). \tag{7}$$

In Eq. (7) m stands for the number of estimated parameters and has to be taken into account while applying a criterion for model selection to avoid an over-fit. This sounds simple and reasonable, but it is not fulfilled in every model (cf. Raschke, 2014). The sample size is again n in Eq. (7), and $\log(L(\theta))$ is the logarithmized likelihood function (cf. Eq. 4). A smaller BIC indicates the better model, the smallest the best. A larger difference between the criteria of alternative models indicates a better differentiation. BIC is very popular in statistics (Lindsey, 1996; Upton and Cook, 2006; Ismail and Jemain, 2007; Claeskens and Hjort, 2008), based on the Kullback–Leibler distance and also related to the likelihood ratio test. In contrast to the latter, the BIC also considers the case of equal number of parameter, and many alternative models can be easily compared. If the BIC selects the models (or the most models) with over-dispersion instead of a model (or the most models) without, then the statistical significance of the over-dispersion is proved.

I do not apply any goodness-of-fit test here because there is no special goodness-of-fit test for the NBD or GPD (cf. Stephend, 1986), and moreover the χ^2 test does not work well for small sample size (see, e.g., Raschke, 2009). The goodness of fit was also not discussed by Karremann et al. (2014).

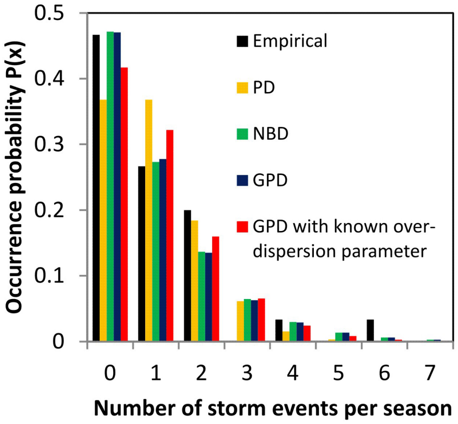

Figure 2. Estimated distributions for the DWD sample with $\mathrm{CRP} = 1$.

5 Over-dispersion of winter storms in Germany

Karremann et al. (2014) have published samples of numbers X of winter storms per season. They have counted storms with a RP of their storm magnitude which fulfill the condition $\mathrm{RP} \geq \mathrm{CRP}$. Therein the cases are presented where $\mathrm{CRP} = 1$, 2 and 5 year. The data are listed in the Supplement and include a sample of historic storms from a period of 30 years ($n = 30$), with magnitudes calculated based on wind station data (DWD) and on data from reanalyzed historic storms (NCEP, ERAI), as well as a very large sample ($n = 4092$) obtained from a climate model simulation ($\mathrm{GCM_{corr}}$). I do not consider the other samples presented by Karremann et al. (2014). The storm magnitudes (return level with return periods in Coles, 2001) are measured by Karremann et al. (2014) by using two indexes which consider the wind speed of different sites (stations or grid points) in Germany and beyond. Details and issues such as the quantification of storm magnitudes are not topics of this work and do not affect the introduced models and methods.

In a first analysis I have estimated for all samples the parameters for distributions PD, NBD and the GPD. The results are listed in Table 1. In every ML estimation applies $\hat{E}(X) = 1/\mathrm{CRP}$, as described in the previous sections. The estimated distributions for DWD data with $\mathrm{CRP} = 1$ year are shown in Fig. 2. It is clear that the NBD and the GPD are very similar. Additionally, the BIC for the NBD and GPD are almost equal for all samples (Table 1). NBD and GPD approximate each other as assumed in Sect. 2. However, the BIC detect the PD as the best model for 56 % of the samples of historic storms. The over-dispersion is detected only in 44 % of the historic storms. But over-dispersion is detected clearly for all $\mathrm{GCM_{corr}}$ samples from climate model simulations.

In a second analysis, I have linked the estimations of different samples. For the large $\mathrm{GCM_{corr}}$ sample of climate model simulations with $\mathrm{CRP} = 1$ year, the estimated over-dispersion parameter of the GPD is $\hat{\beta} = 0.307 \pm 0.031$. I ap-

Table 1. BIC for some samples of winter storm in Germany according to Karremann et al. (2014) ($n = 4092$ for GCM$_{\text{corr}}$, otherwise $n = 30$).

Sample	PD, $m = 1$	NBD, $m = 2$	GPD, $m = 2$	GPD with β =0.307 of GCM$_{\text{corr}}$, $m = 1$
GCM$_{\text{corr}}$, CRP = 1	11 253.55	11 121.61	11 123.50	–
GCM$_{\text{corr}}$, CRP = 2	7800.35	7757.99	7758.58	–
GCM$_{\text{corr}}$, CRP = 5	4432.85	4428.19	4428.15	–
NCEP, CRP = 1	83.04	86.38	86.38	83.40
NCEP, CRP = 2	60.55	63.57	63.58	60.19
NCEP, CRP = 5	38.30	37.94	37.95	37.16
ERAI, CRP = 1	84.43	87.33	87.33	84.01
ERAI, CRP = 2	59.17	62.55	62.55	59.32
ERAI, CRP = 5	36.10	39.24	39.25	35.93
DWD, CRP = 1	91.23	89.76	89.69	87.32
DWD, CRP = 2	66.54	63.43	63.33	63.21
DWD, CRP = 5	38.30	37.94	37.95	37.16

ply this parameter value to estimate the GPD for the historic storm samples. Therein GPD is now parametrized by the expectation and the over-dispersion parameter. The latter is now known and only the expectation is estimated. Both determine the variance $V(X)$ according to Eq. (6). This estimation procedure is possible because the over-dispersion parameter does not depend on the CRP. The GCM$_{\text{corr}}$ sample is relatively large ($n = 4092 \gg 30$) and independent of the samples of historic storms. Furthermore, all samples are from the same random variable X – the number of winter storms per season in Germany for the current climate. Using this special procedure with one known parameter, over-dispersion in the data of historic winter storms in Germany can be properly detected (Table 1, last column). The BIC selects the GPD with known over-dispersion $\beta = 0.307$ in 78 % of the samples of historic storms. Thus the over-dispersion is largely proved for historic storms.

6 Summary

In this communication, I have improved the detection and modeling of the over-dispersion of winter storm occurrence. For this purpose, the GPD, information criterion for the model selection and an over-dispersion parameter have been introduced. The GPD has the advantage of being more universal than the NBD previously used. The application of the information criterion BIC indirectly ensures statistical significance. In addition, the over-dispersion parameter β is more universal than the dispersion statistics ϕ because the over-dispersion parameter is the same for every CRP. All this elements are used in the analysis of the winter storm data for Germany by Karremann et al. (2014). The statistical over-dispersion in historical storm time series for recent decades could be largely proven by a combination of these elements. In this way, the basic conclusions by Karremann et al. (2014) – i.e., that there is over-dispersion in winter storms in Germany – have been confirmed at a higher level of statistical analysis.

The Supplement related to this article is available online at doi:10.5194/nhess-15-1757-2015-supplement.

Acknowledgements. I gratefully acknowledge Joaquim G. Pinto and Matthias Klawa for the fruitful discussions about clustering of storms.

Edited by: B. Merz
Reviewed by: J. G. Pinto and F. Serinaldi

References

Claeskens, G. and Hjort, N. L.: Model selection and model averaging, in: Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2008.

Coles, S.: An introduction to statistical modeling of extreme values, Springer, London, 2001.

Consul, P. C. and Jain, G. C.: A generalization of the Poisson distribution, Technometrics, 15, 791–799, 1973.

Consul, P. C. and Shoukri, M. M.: Maximum likelihood estimation for the generalized Poisson distribution, Commun. Stat. Theory Meth., 10, 977–991, 1984.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B.: Regression – models, methods and applications, Springer, Heidelberg, Germany, 2013.

Ismail, N. and Jemain, A. A.: Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Model, Casualty Actuarial Society Forum, Arlington, VA, 103–158, 2007.

Joe, H. and Zhu, R.: Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution, Biometrical J., 47, 219–229, 2005.

Johnson, N. L., Kemp, A. W., and Kotz, S.: Univariate discrete distributions, in: Wiley Series in Probability and Statistics, 3rd Edn., Wiley, New York, USA, 208–247, 2005.

Karremann, M. K., Pinto, J. G., von Bomhard, P. J., and Klawa, M.: On the clustering of winter storm loss events over Germany, Nat. Hazards Earth Syst. Sci., 14, 2041–2052, doi:10.5194/nhess-14-2041-2014, 2014.

Lindsey, J. K.: Parametric statistical inference, in: Oxford Science Publications, Oxford University Press, Oxford, UK, 1996.

Mack, T.: Schadenversicherungsmathematik (Non-life insurance mathematics), 2nd Edn., Deutsche Gesellschaft für Versicherungsmathematik, Karlsruhe, Germany, 332–334, 2002.

Mailier, P. J., Stephenson, D. B., Ferro, C. A. T., and Hodges, K. I.: Serial Clustering of extratropical cyclones, Mon. Weather Rev., 134, 2224–2240, 2006.

Ogata, Y.: Exploratory analysis of earthquake clusters by likelihood-based trigger models, J. Appl. Probab., 38A, 2002–2012, 2001.

Pinto, J. G., Bellenbaum, N., Karremann, M. K., and Della-Marta, P. M.: Serial clustering of extratropical cyclones over the North Atlantic and Europe under recent and future climate conditions, J. Geophys. Res.-Atmos., 118, 12476–12485, 2013.

Pinto, J. G., Gómara, I., Masato, G., Dacre, H. F., Woollings, T., and Caballero, R.: Large-scale dynamics Associated with clustering of extra-tropical cyclones affecting Western Europe, J. Geophys. Res.-Atmos., 119, 13704–13719, 2014.

Raschke, M.: The Biased Transformation and Its Application in Goodness-of-Fit Tests for the Beta and Gamma Distribution, Commun. Stat. Simul. Comput., 38, 1870–1890, 2009.

Raschke, M.: Insufficient statistical model development of ground motion relations for regions with low seismicity, Bull. Seismol. Soc. Am., 104, 1002–1005, 2014.

Raschke, M. and Thürmer, K.: Diskussion der Modellselektion und weiterer Defizite in der Hochwasserstatistik, Wasser Abfall, 06/2010, 47–51, 2010.

Ross, S. M.: Introduction to probability models, 9th Edn., Elsevier, Amsterdam, the Netherlands, 2007.

Sakamoto, C. M.: Application of the Poisson and Negative Binomial Models to Thunderstorm and Hail Days, Mon. Weather Rev., 101, 350–355, 1973.

Schwarz, G.: Estimating the Dimension of a Model, Ann. Stat., 6, 461–464, 1978.

Stephens, M. A.: Tests based on EDF statistics, in: Goodness-of-Fit Techniques, Statistics: Textbooks and Monographs, Vol. 68, edited by: D'Augustino, R. B. and Stephens, M. A., Marcel Dekker, New York, 97–194, 1986.

Upton, G. and Cook, I.: A dictionary of statistics, 2nd Edn., Oxford University Press, Oxford, UK, 2006.

Vitolo, R., Stephenson, D. B., Cook, I. M., and Mitchell-Wallace, K.: Serial clustering of intense European storms, Meteorol. Z., 18, 411–424, 2009.