

A Bayesian spatio-temporal model for precipitation extremes - STOR team contribution to the EVA2017 challenge

Anna Maria Barlow¹ · Christian Rohrbeck² · Paul Sharkey³ · Rob Shooter¹ · Emma S. Simpson¹

Received: date / Accepted: date

Abstract This paper concerns our approach to the EVA2017 challenge, the aim of which was to predict extreme precipitation quantiles across several sites in the Netherlands. Our approach uses a Bayesian hierarchical structure, which combines Gamma and generalised Pareto distributions. We impose a spatio-temporal structure in the model parameters via an autoregressive prior. Estimates are obtained using Markov chain Monte Carlo techniques and spatial interpolation. This approach has been successful in the context of the challenge, providing reasonable improvements over the benchmark.

Keywords Bayesian hierarchical modelling · Extreme value analysis · Markov chain Monte Carlo · Precipitation extremes · Spatio-temporal dependence

1 Introduction

Recently, there have been numerous examples of devastating rainfall events - these include Storm Desmond, which hit northern England and Scotland, and Hurricane Harvey which affected the southern United States. In both cases, a large amount of damage and disruption was caused by severe flooding. By better understanding the probability of extreme rainfall events occurring, we can prepare more suitably for these potential flood events by adapting infrastructure appropriately.

The challenge data is comprised of precipitation readings for multiple weather stations in the Netherlands; the training set consists of data collected between 1972 and 1995 whilst the validation set was collected from 1996 to

¹ STOR-i Centre for Doctoral Training, Department of Mathematics and Statistics, Lancaster University, LA1 4YR, UK

² Data Science Institute, Department of Mathematics and Statistics, Lancaster University, LA1 4YF, UK

³ JBA Consulting, 4th floor, 1 Belle Vue Square, Broughton Road, Skipton, BD23 1FJ, UK

2016, with different numbers of observations for each site. A detailed description of the data is provided in Wintenberger (2018). The aim of the competition is to predict extreme quantiles for the years 1996 to 2016 and predictions are assessed via a predefined error metric; see Wintenberger (2018).

There exists a rich literature within the extreme value theory framework for modelling precipitation extremes. A classical approach is to utilise block maxima. Suppose that we have independent and identically distributed (i.i.d.) random variables X_1, \dots, X_n , with $M_n = \max\{X_1, \dots, X_n\}$. When normalised appropriately, and as $n \rightarrow \infty$, M_n follows a generalised extreme value (GEV) distribution (Fisher and Tippett, 1928), which has distribution function

$$F(x) = \exp \left\{ - \left[1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right]_+^{-\frac{1}{\xi}} \right\}, \quad (1)$$

where $\{z\}_+ = \max\{0, z\}$, and has parameters $(\mu, \sigma, \xi) \in \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}$, corresponding to location, scale and shape parameters respectively.

An alternative technique is to follow Pickands (1975) and use exceedances of a threshold u . For some suitably large u , the conditional distribution function of $(X_i - u) \mid (X_i > u)$ is approximately given by the generalised Pareto distribution (GPD), which has the form

$$H(x) = 1 - \left(1 + \frac{\xi x}{\psi} \right)_+^{-\frac{1}{\xi}}, \quad x > 0, \quad (2)$$

where $(\psi, \xi) \in \mathbb{R}_+ \times \mathbb{R}$ are the scale and shape parameters respectively. In the context of the challenge at hand, both the GEV and GPD may be fitted separately at each site to give a model fit whereby any dependence is ignored.

By considering the physical process of rainfall, one can expect that nearby locations will exhibit similar behaviour, which invites improved inference by sharing information across sites. One popular method for the modelling of spatial extremes is to use max-stable processes (Brown and Resnick, 1977; Smith, 1990; Schlather, 2002). These arise as the limiting process from replications of spatial processes which have been suitably normalised (de Haan, 1984) and have been used to analyse rainfall data previously; see, for example, Davison et al (2012) and Reich and Shaby (2012). However, such processes assume dependence of the extremes across sites; an investigation of pairwise dependence using scatter plots showed no clear evidence for this behaviour across the spatial grid. Moreover, max-stable models are difficult to fit and this would have been further impeded by the lack of data available at some sites.

Another approach is to impose spatial structure on the model parameters via a Bayesian hierarchical model; this is closer in nature to the method we propose. Spatial hierarchical models have been used previously to model spatial count data (Diggle et al, 1998) and, more recently, have been utilised in extreme value analysis. Cooley et al (2007) describe a model, applied to rainfall data, whereby a GPD is fitted at the sampling locations, and allow the model parameters to vary according to a spatial process structure - in particular the authors use a Gaussian process for this. A spatio-temporal hierarchical

modelling method for extreme events is given by Sang and Gelfand (2009), who apply their methods to precipitation data.

In this paper, we define a Bayesian hierarchical model which accounts for the spatial and seasonal variation in the data. Our approach captures the frequency of non-zero events of precipitation and introduces an extremal mixture model, combining Gamma and generalised Pareto distributions, for positive amounts of rainfall. Spatio-temporal structure in the parameters for the extremal mixture model is imposed via a separate autoregressive prior for each of them, which takes the form of a Gaussian Markov random field. Model estimates are then obtained using spatial interpolation and Markov chain Monte Carlo (MCMC) techniques. Cooley et al (2007) defines a similar approach for continuous space, whereas we consider a finite number of sites and additionally incorporate seasonality.

The remainder of this article is structured as follows. Section 2 details our Bayesian framework and its estimation: in Sections 2.1 and 2.2 respectively, we specify our likelihood and prior models; in Section 2.3, we discuss parameter estimation. In Section 3, we discuss the results obtained using our method for modelling rainfall extremes, and highlight areas for potential improvements.

2 Methodology

2.1 Likelihood

Interest lies in modelling the daily rainfall amounts for each site and month. Due to seasonality in the rainfall data, the weak extremal dependence of the daily amount of rainfall across sites and the nature of the challenge, we model each month and site individually. Specifically, daily rainfall events within a month at a site are assumed to be i.i.d. Our model is motivated by an analysis of the sites for which data have been recorded for at least five years.

Let $R_{j,m}$ denote the random variable corresponding to the daily rainfall amount at site j for a day in month $m = 1, \dots, 12$. We consider the transformed random variable

$$\tilde{R}_{j,m} = \log(1 + R_{j,m}). \quad (3)$$

Wadsworth et al (2010) show that such a transformation may increase the rate of convergence of the distribution tails to an extreme value form, in particular for distributions which appear as heavy-tailed as our rainfall data. Predictions on the extreme quantiles of $R_{j,m}$ are later obtained in Section 3 by reversing this transformation. We note that the transformed observations are non-negative and an observation of $R_{j,m} = 0$ remains unchanged.

We infer on the distribution of $\tilde{R}_{j,m}$ by defining a hierarchical model. The first model component considers occurrences of non-zero amounts of rainfall on a day, $\tilde{R}_{j,m} > 0$, and we denote their probability by $p_{j,m}$. A temporal trend in $p_{j,m}$ was investigated, but we did not find evidence of this for any site. Next, we consider the distribution $\tilde{R}_{j,m} | (\tilde{R}_{j,m} > 0)$. There exists a rich literature on modelling positive rainfall amounts, such as Wilks (2006), So

et al (2015) and Yunus et al (2017). By investigating QQ plots, we find that an estimated Gamma distribution works quite well for non-extreme amounts of precipitation. However, most of the observed monthly extremes are not captured well.

To improve the model fit, we define an extremal mixture model (Frigessi et al, 2002; Behrens et al, 2004; MacDonald et al, 2011) which combines the Gamma distribution with a GPD as defined in (2). Given a threshold $u_{j,m}$, $\tilde{R}_{j,m} \mid (\tilde{R}_{j,m} \leq u_{j,m})$ follows a truncated Gamma distribution, while $\tilde{R}_{j,m} \mid (\tilde{R}_{j,m} > u_{j,m})$ is generalised Pareto distributed. Formally, let $G_{j,m} \sim \text{Gamma}(\alpha_{j,m}, \beta_{j,m})$ with shape $\alpha_{j,m}$ and rate $\beta_{j,m}$, and $H_{j,m} \sim \text{GPD}(\psi_{j,m}, \xi_{j,m})$ with scale $\psi_{j,m} = \tilde{\psi}_{j,m} - \xi u_{j,m}$ and shape $\xi_{j,m}$. The reparametrisation of the scale parameter in $H_{j,m}$ removes the effect of the threshold on inference and has been used in previous studies (Fawcett and Walshaw, 2006). Then, the cumulative distribution function of $\tilde{R}_{j,m} \mid (\tilde{R}_{j,m} > 0)$ is given by

$$\mathbb{P}(\tilde{R}_{j,m} > r \mid \tilde{R}_{j,m} > 0) = \begin{cases} \mathbb{P}(G_{j,m} > r) & r \leq u_{j,m}, \\ \mathbb{P}(G_{j,m} > u_{j,m}) \mathbb{P}(H_{j,m} > r - u_{j,m}) & r > u_{j,m}. \end{cases} \quad (4)$$

Combining the model components defined above, the event $\tilde{R}_{j,m} > r$, for $r > u_{j,m}$, occurs with probability

$$\mathbb{P}(\tilde{R}_{j,m} > r) = p_{j,m} \mathbb{P}(G_{j,m} > u_{j,m}) \mathbb{P}(H_{j,m} > r - u_{j,m}).$$

Due to the empirical mean of $R_{j,m} \mid (R_{j,m} > 0)$ being similar for all j , we fix $\alpha_{j,m}$, $m = 1, \dots, 12$, in the Gamma distribution to be constant across sites and, thus, refer to this parameter as α_m in the rest of this paper.

2.2 Prior model

Prior selection is critical in this analysis due to the varying degrees of data availability at each site; inference at sites where data are lacking or unavailable will be dominated by the prior distribution. We considered uninformative, improper Uniform priors on $\log \alpha_m$, $\log \beta_{j,m}$, $\log \tilde{\psi}_{j,m}$ and $\xi_{j,m}$. However, these produced unrealistic estimates of $\xi_{j,m}$, mostly due to the difficulty in estimating $\xi_{j,m}$ given short data records. Studies on extreme rainfall often feature the prior used in Martins et al (2000) which constrains the shape parameter to be in a sensible interval.

We instead introduce a prior aimed at exploiting the spatial and seasonal structure of the model parameters. We assume that parameters for neighbouring sites and adjacent months are likely to be similar. Explicitly, we propose for $\phi_{j,m}$, an arbitrary parameter at site j and month m , that

$$\phi_{j,m} \sim \mathcal{N}\left(\frac{\phi_{j,m-1} + \phi_{j,m+1} + \sum_{j' \neq j} \phi_{j',m} d_{j,j'}}{2 + \sum_{j' \neq j} d_{j,j'}}, \frac{1}{(2 + \sum_{j' \neq j} d_{j,j'}) \tau_\phi}\right), \quad (5)$$

where $\tau_\phi > 0$ denotes the precision for parameter ϕ , common to all sites and months. The constant $d_{j,j'} \geq 0$ describes our prior belief concerning the degree of similarity of $\phi_{j,m}$ and $\phi_{j',m}$. This prior is a variant of the Intrinsic Autoregressive (IAR) prior as described in Banerjee et al (2004) and allows us to pool information across neighbouring sites and months, which helps to produce more stable parameter estimates and to reduce uncertainty in these estimates. The cyclical nature of the sequence of months means that values 0 and 13 for $m-1$ and $m+1$ should be replaced by the values 12 and 1 respectively in order to ensure that December and January are correctly identified as being adjacent months. We define a flat, conjugate Gamma(1, 0.001) prior for τ_ϕ .

2.3 Threshold selection and estimation

We detail our approach to estimate the model defined in Sections 2.1 and 2.2 in the following. First, we consider $p_{j,m}$, which can be estimated independently from the remaining parameters due to the hierarchical model structure. Next, the selection of the thresholds $u_{j,m}$ is described. Finally, we infer on the remaining model parameters via an MCMC algorithm which is outlined at the end of this subsection.

For sites with more than five years of data, we estimate $p_{j,m}$ empirically due to the high number of observations available. We infer on the remaining sites via spatial interpolation. Let \mathcal{J} denote the indices of the sites with at least five years of data. We further define a pairwise weighting between arbitrary sites j and j' by introducing the weight

$$d_{j,j'} = \exp(-\|\mathbf{x}_j - \mathbf{x}_{j'}\|), \quad (6)$$

where \mathbf{x}_j denotes the longitude and latitude coordinates of site j and $\|\cdot\|$ corresponds to the Euclidean distance. As the study region is small, the curvature of the earth is negligible and the Euclidean distance in the two-dimensional space is close to the true distance between the sites. Then for a site $j \notin \mathcal{J}$, the estimate $\hat{p}_{j,m}$ for $p_{j,m}$ is derived as

$$\hat{p}_{j,m} = \sum_{j' \in \mathcal{J}} d_{j,j'} \hat{p}_{j',m}. \quad (7)$$

The weights $d_{j,j'}$ defined in (6) are identical to the ones which we set in the prior density (5). As the weighting function (6) produces larger values for locations close together, a higher weight is given to neighbouring sites.

We now consider how to select the thresholds, $u_{j,m}$, of our model (4). These thresholds must be large enough for the asymptotic argument of Pickands (1975) to approximately hold whilst also low enough so that we have a sufficient number of observations for reliable model fitting. We use the classical fixed threshold approach as described in Coles (2001) for the sites in \mathcal{J} . Specifically, by inspection of threshold stability plots, we find the smallest threshold

above which the GPD is an appropriate model for the exceedances. For the other sites, we estimate these thresholds in an equivalent manner to (7). Other threshold selection methods are outlined by Scarrott and MacDonald (2012).

The parameters of our Gamma-GPD mixture model are estimated using MCMC methods. We sample from the posterior distribution using a Metropolis-within-Gibbs scheme. In particular, proposal values of each parameter are generated sequentially from a Gaussian distribution and accepted with a probability defined as the posterior ratio of the proposed state relative to the current state of the Markov chain. The hyperparameter τ_ϕ in (5) is updated by sampling from the full conditional Gamma posterior as described by Knorr-Held (2003). We tune the parameters of the MCMC algorithm to ensure an acceptance rate of 20-25% in accordance with the optimality criterion of Roberts et al (1997).

3 Results and Discussion

We begin this section by considering the results of the MCMC implementation. We run our MCMC chains for 20000 iterations, and discard the first 5000 iterations as burn-in to aid convergence. Examples of the chains produced are provided in Figure 1 for scale and shape parameters $\psi_{10,6}$ and $\xi_{10,6}$. Estimates of these parameters were obtained using the posterior means of their respective MCMC chains. These plots demonstrate that good mixing has been achieved for this case; similar results were obtained across other stations and months.

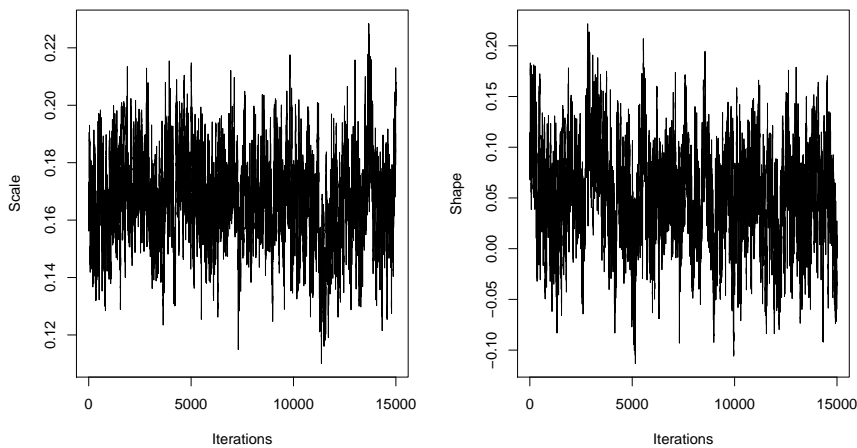


Fig. 1: MCMC chains for the scale and shape parameters for station 10 in June.

We now explore the monthly variation in the estimated model parameters by focussing on results at four nearby stations. The locations of these stations are shown in the top left panel of Figure 2. The data set contains over 8000 observations for stations 2 and 5, and no observations for stations 7 and 10. The top right and bottom left panels of Figure 2 show our estimates of the scale and shape parameters, respectively, at these four locations. These plots demonstrate the seasonality in the parameter estimates, with higher values of both the scale and shape generally corresponding to summer and autumn months. This effect is maintained in the predicted 0.998 quantiles, shown in the bottom right panel of Figure 2, which are typically highest between June and October. A similar trend was observed at other sites, particularly those with limited data where estimates are more heavily influenced by information from other locations, due to the spatial smoothing imposed by the model.

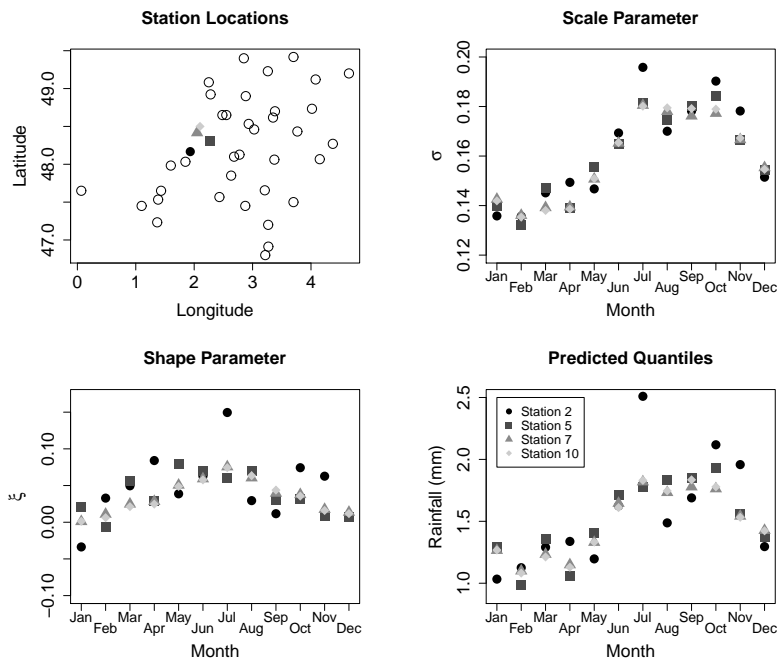


Fig. 2: Location of stations 2, 5, 7 and 10, as well as estimates of the corresponding scale and shape parameters and predicted 0.998 quantiles.

We now consider our estimates in the context of the competition, which used the quantile loss function by Koenker (2005). In particular, as in the challenge, we consider the percentage improvement provided by our method over benchmark predictions. The competition was split into two challenges: Challenge 1 involved only sites where observations were available, with the

benchmark quantile estimates being given by the monthly maxima at each station; Challenge 2 included predictions for all sites, with the benchmark for those sites with no data being taken as the average of the quantiles predicted in Challenge 1 for each month. Our method gave a 59.9% improvement over the benchmark for Challenge 1, and a 57.7% improvement for Challenge 2. Table 1 shows the performance of our approach using this same metric, but with the results separated by month.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Challenge 1	57.7	71.1	60.0	65.0	43.7	62.8	65.9	77.0	38.7	38.4	52.2	33.4
Challenge 2	54.4	69.3	57.4	61.9	43.1	60.7	64.2	75.4	37.9	36.4	49.3	31.3

Table 1: Percentage improvement over the benchmark for Challenges 1 and 2 across each month.

As is to be expected, our method performed better in Challenge 1, where only predictions for sites with observations were considered, across all months. Looking at these results separately for each month allows us to identify possible areas for improvement. In particular, the scores for September, October and December are lower than for other months, suggesting that the method could be improved by focussing on the modelling of autumn and winter months.

Compliance with Ethical Standards

Conflict of Interest: The authors declare that they have no conflict of interest.

Acknowledgements The authors gratefully acknowledge the support of the Engineering and Physical Sciences Research Council (EPSRC) funded STOR-i Centre for Doctoral Training (Grant numbers EP/H023151/1 and EP/L015692/1). Rohrbeck also acknowledges funding by the Faculty of Science and Technology, Lancaster University. The authors also acknowledge sponsorship of EDF Energy (P. Sharkey), Shell Research (R. Shooter), JBA Risk Management (A. Barlow) and the Joint UK BEIS/Defra Met Office Hadley Centre Climate Programme (GA01101) (P. Sharkey). The authors also extend their thanks to Jonathan Tawn and the referees for helpful comments.

References

- Banerjee S, Carlin BP, Gelfand AE (2004) Hierarchical Modeling and Analysis for Spatial Data. CRC Press, Boca Raton
- Behrens CN, Lopes HF, Gamerman D (2004) Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling* 4(3):227–244
- Brown BM, Resnick SI (1977) Extreme values of independent stochastic processes. *Journal of Applied Probability* 14(4):732–739
- Coles SG (2001) An introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics, Springer-Verlag London

- Cooley D, Nychka D, Naveau P (2007) Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association* 102(479):824–840
- Davison AC, Padoan SA, Ribatet M (2012) Statistical modeling of spatial extremes. *Statistical Science* 27(2):161–186
- Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47(3):299–350
- Fawcett L, Walshaw D (2006) A hierarchical model for extreme wind speeds. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 55(5):631–646
- Fisher RA, Tippett LHC (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society* 24(2):180–190
- Frigessi A, Haug O, Rue H (2002) A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes* 5(3):219–235
- de Haan L (1984) A spectral representation for max-stable processes. *The Annals of Probability* 12(4):1194–1204
- Knorr-Held L (2003) Some remarks on Gaussian Markov random field models for disease mapping. In: Green PJ, Hjort NL, Richardson S (eds) *Highly Structured Stochastic Systems*, Oxford University Press, Oxford, pp 203–207
- Koenker R (2005) *Quantile Regression*. Cambridge University Press, Cambridge
- MacDonald A, Scarrott C, Lee D, Darlow B, Reale M, Russell G (2011) A flexible extreme value mixture model. *Computational Statistics & Data Analysis* 55(6):2137–2157
- Martinez ES, Stedinger JR, et al (2000) Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research* 36(3):737–744
- Pickands J (1975) Statistical inference using extreme order statistics. *The Annals of Statistics* 3(1):119–131
- Reich BJ, Shaby BA (2012) A hierarchical max-stable spatial model for extreme precipitation. *The Annals of Applied Statistics* 6(4):1430–1451
- Roberts GO, Gelman A, Gilks WR (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability* 7(1):110–120
- Sang H, Gelfand AE (2009) Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics* 16(3):407–426
- Scarrott C, MacDonald A (2012) A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT Statistical Journal* 10(1):33–60
- Schlather M (2002) Models for stationary max-stable random fields. *Extremes* 5(1):33–44
- Smith RL (1990) Max-stable processes and spatial extremes. Unpublished manuscript

- So BJ, Kwon HH, Kim D, Lee SO (2015) Modeling of daily rainfall sequence and extremes based on a semiparametric Pareto tail approach at multiple locations. *Journal of Hydrology* 529:1442–1450
- Wadsworth JL, Tawn JA, Jonathan P (2010) Accounting for choice of measurement scale in extreme value modeling. *The Annals of Applied Statistics* 4(3):1558–1578
- Wilks DS (2006) *Statistical Methods in the Atmospheric Sciences*, 2nd edn. International Geophysics Series ; 59, Elsevier, Amsterdam
- Wintenberger O (2018) Editorial: special issue on the Extreme Value Analysis conference challenge ‘Prediction of extremal precipitation’. *Extremes* (To Appear)
- Yunus RM, Hasan MM, Razak NA, Zubairi YZ, Dunn PK (2017) Modelling daily rainfall with climatological predictors: Poisson-Gamma generalized linear modelling approach. *International Journal of Climatology* 37(3):1391–1399