2021-04-05

Social Media Integration of Flood Data: A Vine Copula-Based Approach

Ansell, L

http://hdl.handle.net/10026.1/17049

Arxiv.org

Social Media Integration of Flood Data: A Vine Copula-Based Approach

Lauren Ansell and Luciana Dalla Valle

University of Plymouth

April 6, 2021

Abstract

Floods are the most common and among the most severe natural disasters in many countries around the world. As global warming continues to exacerbate sea level rise and extreme weather, governmental authorities and environmental agencies are facing the pressing need of timely and accurate evaluations and predictions of flood risks. Current flood forecasts are generally based on historical measurements of environmental variables at monitoring stations. In recent years, in addition to traditional data sources, large amounts of information related to floods have been made available via social media. Members of the public are constantly and promptly posting information and updates on local environmental phenomena on social media platforms. Despite the growing interest of scholars towards the usage of online data during natural disasters, the majority of studies focus exclusively on social media as a stand-alone data source, while its joint use with other type of information is still unexplored. In this paper we propose to fill this gap by integrating traditional historical information on floods with data extracted by Twitter and Google Trends. Our methodology is based on vine copulas, that allow us to capture the dependence structure among the marginals, which are modelled via appropriate time series methods, in a very flexible way. We apply our methodology to data related to three different coastal locations in the South cost of the UK. The results show that our approach, based on the integration of social media data, outperforms traditional methods, providing a more accurate evaluation and prediction of flood events.

Keywords: Dependence Modelling; Vine Copulas; Time Series Modelling; Social Media Sentiment Analysis; Floods; Natural Hazards; Climate Change.

1 Introduction

In recent years, climate change has caused an exacerbation of the frequency and severity of natural hazard phenomena, such as floods, storms, wildfires and other extreme weather events (Field et al., 2012; Muller et al., 2015). Around the world, a substantial part of the population is exposed to flood risk, with more than 2.3 billion people - which is about one third of the world's population - who live in locations that are estimated to experience some level of inundation during flood events (UN, 2015). In the United Kingdom, intense storms occurred during recent years, bringing severe flooding and causing considerable damage to people, infrastructure and the economy, totalling millions of pounds (Smith et al., 2017). This caused a growing need for timely and accurate information about the severity of flooding, which is essential for forecasting and nowcasting these phenomena and for effectively managing response operations and appropriately allocate resources (Rosser et al., 2017).

Generally, statistical and machine learning models are employed to estimate and predict inundations and, typically, the information used is gathered from meteorological and climatological instrumentation at monitoring stations. For example, Wang and Du (2003) use a combination of meteorological, geographical and urban data to produce flooding tables and maps published via Internet for public consultation. Keef et al. (2013) used data from a set of UK river flow gauges to estimate the probability of widespread floods based on the conditional exceedance model of Heffernan and Tawn. Grego et al. (2015) collected historic flood frequency data and modelled them via finite mixture models of stationary distributions using censored data methods. Balogun et al. (2020) utilized geographic information system and remote sensing data from Malaysia to generate flood susceptibility maps, applying Fuzzy-Analytic Network Process flood models. Moishin et al. (2020) investigated fluvial flood risk in Fiji developing a flood index based on current and antecedent day's precipitation. Talukdar et al. (2020) gathered historical flood data related to the Teesta River basin in Bangladesh and employed ensemble machine learning algorithms to predict flooding sites and flood susceptible zones.

However, information collected at monitoring stations may suffer from data sparsity, time delays and high costs (Muller et al., 2015). In particular, remotely sensed data may take several hours to become available (Mason et al., 2012) and their temporal resolution is often limited (Schumann et al., 2009).

On the other hand, an increasing availability of consumer devices, such as smartphones and tablets, is leading to the dissemination and communication of flood events directly by individuals, with information shared in real-time using social media. User-generated content shared online often includes reports on

meteorological conditions especially in case of extreme or unusual weather (Alam et al., 2018). Recent studies focused specifically on social media sources, such as Twitter, Facebook and Flickr, to collect real-time information on floods and environmental events and their impacts across the globe. For example, Herfort et al. (2014) and De Albuquerque et al. (2015) identified spatial patterns in the occurrence of flood-related Tweets associated with proximity and severity of the River Elbe flood in Germany in June 2013. Saravanou et al. (2015) performed a case study on the floods that occurred in the UK during January 2014, investigating how these were reflected on Twitter. Twitter data generated during flooding crisis was also used by Spielhofer et al. (2016) to evaluate techniques to be adopted in real-time to provide actionable intelligence to emergency services. Different methods to create flood maps from Twitter micro-blogging were presented by Brouwer et al. (2017), Smith et al. (2017) and Arthur et al. (2018), who applied their approaches to different locations, such as the city of York (UK), Newcastle upon Tyne (UK) and the whole England region, respectively. The 2015 South Carolina flood disaster was analysed by Li et al. (2018) to map the flood in real time by leveraging Twitter data in geospatial processes. Spruce et al. (2021) analysed rainfall events occurred across the globe in 2017, comparing outputs from social sensing against a manually curated database created by the Met Office.

However, the majority of contributions in the literature analysing online generated data are focusing exclusively on social media sources, overlooking any relation or synergy with other sources of information. One of the few exceptions is the paper by Rosser et al. (2017), who estimated the flood inundation extent in Oxford (UK) in 2014 based on the fusion of remote sensing, social media and topographic data sources, using a simple Weights-of-evidence analysis.

In this paper we propose to leverage the association between social media and traditional information via sophisticated statistical modelling based on vine copulas, to enhance the assessment and prediction of flood phenomena.

Copulas are multivariate statistical tools, which allow us to model separately the marginal models and their dependence structure (Huang et al., 2017). Copulas were used in flood risk analysis, for example, by Jane et al. (2016) to predict the wave height at a given location by exploiting the spatial dependence of the wave height at nearby locations. The use of copulas in flood risk management was also explored by Jane et al. (2018), who used a copula to capture dependencies in a three dimensional loading parameter space, estimating the overall failure probability. Copulas were also employed by Feng et al. (2020), who used time-varying copulas with nonstationary marginal distributions to estimate the dependence structure of inundation magnitudes in flood coincidence risk assessment.

Vine copulas are based on bivariate copulas as building blocks and provide a great deal of flexibility, compared to standard copulas and other traditional multivariate approaches, in modeling complex dependence structures between the variables. Vine copulas were adopted, for example, by Latif and Mustafa (2020) to simultaneously model trivariate flood characteristics for the Kelantan River basin in Malaysia. Tosunoglu et al. (2020) applied vine copulas in hydrology for multivariate modelling of flood characteristics in the Euphrates River Basin, Turkey.

However, to the best of our knowledge, there are currently no studies exploring the use of vine copulas to integrate social media data with other types of information. This paper proposes a novel approach, based on vine copulas, that combines data gathered from Twitter and Google Trends with remotely-sensed information. The application of our methodology to three different coastal locations in the South of the UK shows that our approach performs better than traditional approaches to estimate and predict flood events.

The remainder of the paper is organised as follows. Section 2 describes the environmental and social media data used in the analysis; Section 3 illustrates the vine copula methodology; Section 4 reports the results of the analysis; finally, concluding remarks are presented in Section 5.

2 Study Area and Data Collection

The UK coastline has been subject to terrible floods throughout history. Over the last few years, storms and floods relentlessly hit the UK coast, triggering intense media coverage and public attention. Table 1 lists the major winter storm events affecting the UK between 2012 and 2018.

In this paper we consider three locations in the South coast of the UK, which were subjected to intense flood events in recent years: Portsmouth, Plymouth and Dawlish. The inundation episodes of the last few years had a substantial socio-economic impact on the local communities of the three locations, which are totalling a population of almost 500,000. The three areas were affected by most of the inundation events listed in Table 1. In particular, devastating overnight storms on February 4, 2014 swept the main rail route at Dawlish, leaving tracks dangling in mid-air. Moreover, on the night of February 14 huge waves damaged a line of shipping containers forming a breakwater at Dawlish and punched a new hole in the sea wall.

In order to accurately estimate and predict flood phenomena in the three coastal areas, we applied the vine copula methodology to data based on historical measurement in conjunction with information gathered online.

Table 1: Major winter storm events in the UK between 2012 and 2018. Note that the storm naming system was introduced in 2015.

Winter 2012/13	**Winter 2013/14**	**Winter 2015/16**		**Winter 2016/17**		**Winter 2017/18**	
Date	**Date**	**Storm Name**	**Date**	**Storm Name**	**Date**	**Storm Name**	**Date**
11 Oct	28 Oct	Abigail	12-13 Nov	Angus	20 Nov	Aileen	12-13 Sep
18 Nov	5-6 Dec	Barney	17-18 Nov	Barbara	23-24 Dec	Brian	21 Oct
14 Dec	18-19 Dec	Clodagh	29 Nov	Conor	25- 26 Dec	Caroline	7 Dec
19 Dec	23-24 Dec	Desmond	5-6 Dec	Doris	23 Feb	Dylan	30-31 Dec
22 Dec	26-27 Dec	Eva	24 Dec	Ewan	26 Feb	Eleanor	2-3 Jan
	30-31 Dec	Frank	29-30 Dec			Fionn	16 Jan
	3 Jan	Gertrude	29 Jan			Georgina	24 Jan
	25-26 Jan	Henry	1-2 Feb				
	31 Jan-1 Feb	Imogen	8 Feb				
	4-5 Feb	Jake	2 Mar				
	8-9 Feb	Katie	27-28 Mar				
	12 Feb						
	14-15 Feb						

For each one of the three locations, we collected daily hydraulic loading condition data as well as social media information for the period between January 2012 and December 2016, obtaining $1,827$ daily data points for each variable. More precisely, we downloaded wave height (m) and water level (tidal residual, m) data from the UK Environment Agency flood-monitoring API [1]. Furthermore, for the aforementioned locations, we gathered Google Trends information on the number of searches for the keywords *flood*, *flooding*, *rain* and *storm*, using the `gtrends` package from the R software (Massicotte and Eddelbuettel, 2021; R Core Team, 2020). In addition, we collected Twitter messages containing the same keywords used to perform Google Trends searches for the three areas. After removing tweets sent by automated accounts, which contained factual information about the current weather in the required location, we obtained 9,781 tweets for Portsmouth, 4,995 tweets for Plymouth and 1,769 tweets for Dawlish. From the Twitter data, we considered the total number of tweets as well as the sentiment scores calculated using two different lexicons: Bing and Afinn (Hu and Liu, 2004), which are available in the R `tidytext` package (Silge and Robinson, 2016). The Bing lexicon splits words into positive or negative. The Bing sentiment score for each tweet is calculated by counting the number of positive words used in each tweet and subtracting from this the number of negative words. The Afinn lexicon scores words between ± 5. The Afinn sentiment score is calculated by multiplying the score of each word by the number of times it appears in the tweet; these scores are then summed to derive the overall sentiment

[1] Available at the website `https://environment.data.gov.uk/flood-monitoring/doc/reference`

score. In order to take into account of the different population sizes living in the three areas [2], we scaled the Bing and Afinn sentiment scores by the relevant number of residents.

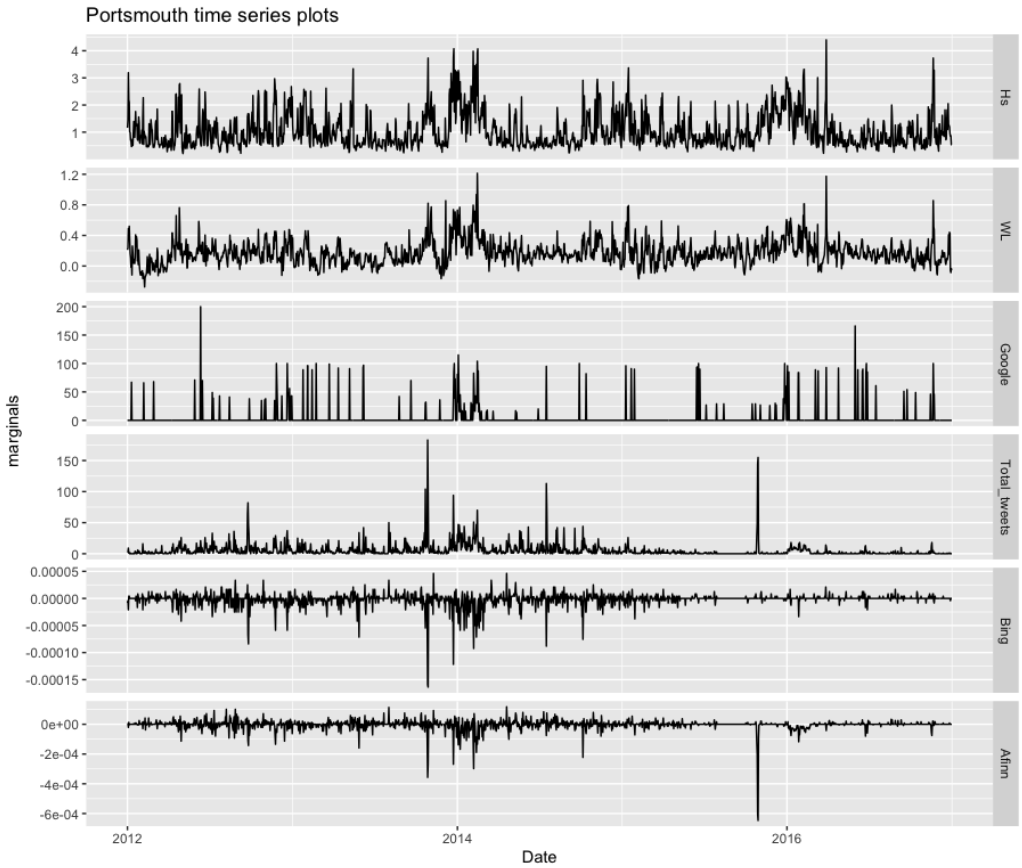

Figure 1: Trace plots of Portsmouth data.

Figures 1, 2 and 3 show the trace plots of the data collected for Portsmouth, Plymouth and Dawlish, respectively. The panels (from top to bottom) illustrate the wave height (`Hs`), the water level (`WL`), the Google Trends searches (`Google`), the total number of Tweets (`Total_tweets`), the Bing sentiment scores (`Bing`) and the Afinn sentiment scores (`Afinn`). We notice spikes in the plots corresponding to most of the storm events listed in Table 1. For example, the flood events occurred

[2]We considered a total population of 238,137 for Portsmouth; a total population of 234,982 for Plymouth; a total population of 16,298 for Dawlish. *Source*: 2011 United Nations population figure, available at: `https://unstats.un.org/unsd/demographic-social/`

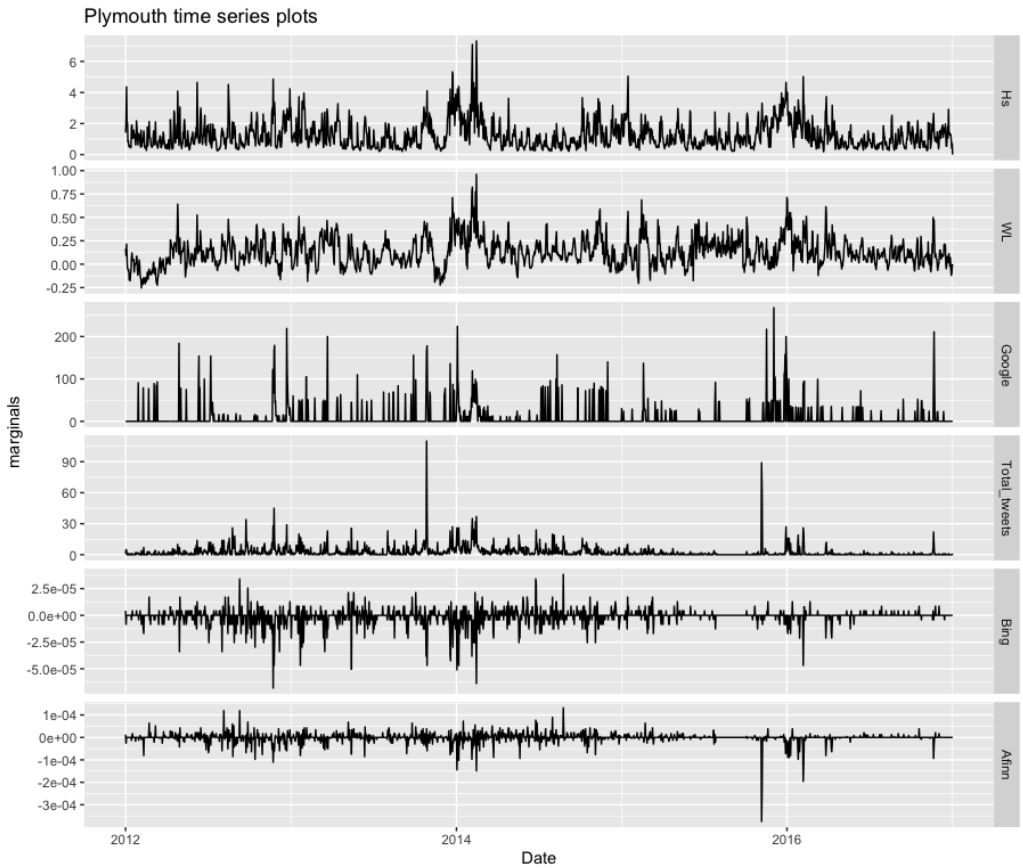

Figure 2: Trace plots of Plymouth data.

in February 2014 are reflected in high spikes in the time series plots, especially for Dawlish in Figure 3. From the plots we also notice that the time series exhibit a similar pattern at specific time points. Generally, the higher the values of wave height and water level, the higher the volume of tweets and Google searches, and the lower the sentiment scores for both lexicons. This suggest the presence of association between the social media and remotely-sensed data.

3 Methodology

The copula is a function that allows us to bind together a set of marginals, to model their dependence structure and to obtain the joint multivariate distribution (Joe, 1997; Nelsen, 2007). Sklar's theorem (Sklar, 1959) is the most important result in

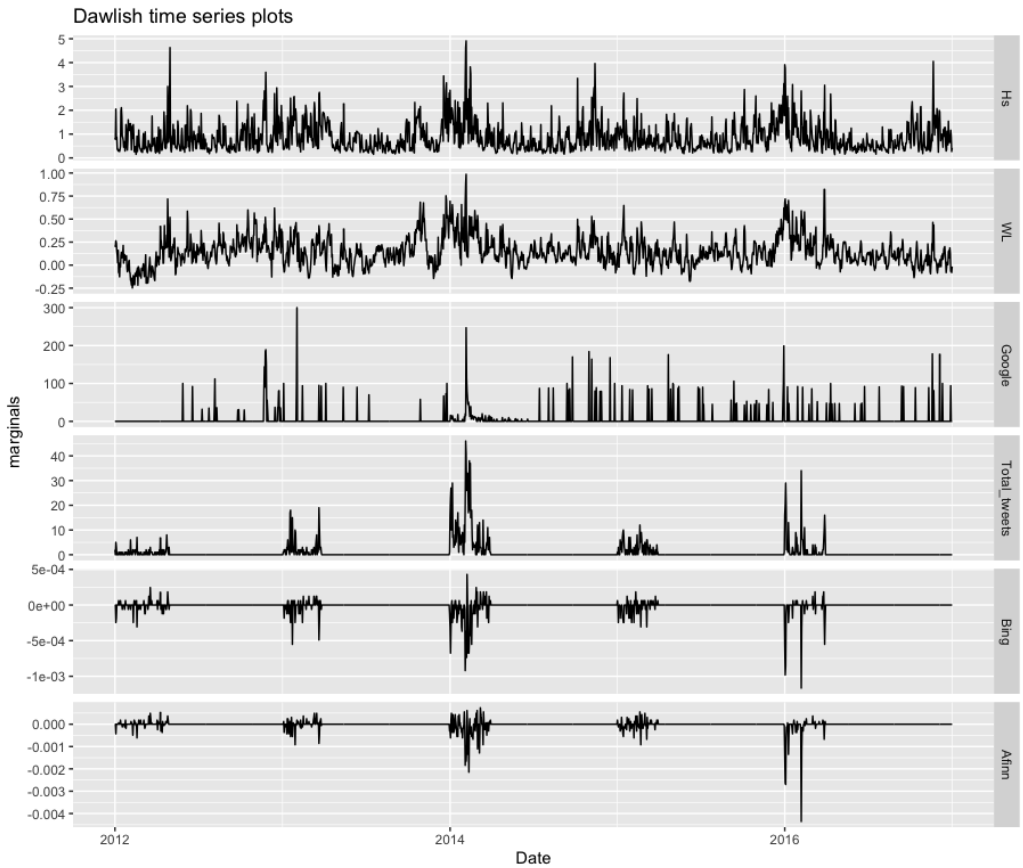

Figure 3: Trace plots of Dawlish data.

copula theory. It states that, given a vector of random variables $\mathbf{X} = (X_1, \ldots, X_d)$, with d-dimensional joint cumulative distribution function $F(x_1, \ldots, x_d)$ and marginal cumulative distributions (cdf) $F_j(x_j)$, with $j = 1, \ldots, d$, a d-dimensional copula C exists, such that

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d); \boldsymbol{\theta}),$$

where $F_j(x_j) = u_j$, with $u_j \in [0, 1]$ are called *u-data*, and $\boldsymbol{\theta}$ denotes the set of parameters of the copula. The joint density function can be derived as

$$f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d); \boldsymbol{\theta}) \cdot f_1(x_1) \cdots f_d(x_d),$$

where c denotes the d-variate copula density. The copula allows us not only to determine the joint multivariate distribution, but also to describe the dependencies

among the marginals, that can potentially be all different and can be modelled using distinct distributions.

In this paper, we adopt the 2-steps inference function for margins (IFM) approach (Joe and Xu, 1996), estimating the marginals in the first step, and then the copula, given the marginals, in the second step.

3.1 Marginal Models

Given the different characteristics of the six marginals, we fitted different models for each of the six time series for each location. Further, we extracted the residuals ε_j, with $j = 1, \ldots, d$, from each marginal model and we applied the relevant distribution functions to get the *u-data* $F_j(\varepsilon_j) = u_j$ to be plugged into the copula.

3.1.1 Wave height (`Hs`)

The best fitting model for the log-transformed `Hs` marginal for all three locations was the autoregressive integrated moving average (ARIMA) model (for more information about ARIMA models, see, for example Hyndman and Athanasopoulos (2018)). The ARIMA model aims to describe the autocorrelations in the data by combining autoregressive and moving average models. The model is usually denoted as ARIMA(p, d, q), where the values in the brackets indicate the parameters: p, d, q, where p is the order of the autoregressive part, d is the degree of first differencing involved and q is the order of the moving average part. The ARIMA model, for $t = 1, \ldots, T$ takes the following form:

$$y_t = a + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{1}$$

where $y_t = (1 - B)^d x_t$, x_t are the original data values, B is the backshift operator, a is a constant, ϕ_i, with $i = 1, \ldots, p$, are the autoregressive parameters, θ_i, with $i = 1, \ldots, q$, are the moving average parameters and $\varepsilon_t \sim N(0, 1)$ is the error term.

3.1.2 Water level (`WL`)

We fitted the log-transformed `WL` marginal for the Plymouth location with an ARIMA model, as described in Eq.(1). However, for Portsmouth and Dawlish, the ARIMA-GARCH model with Student's t innovations appeared to have a better fit. This model combines the features of the ARIMA model with the generalized autoregressive conditional heteroskedastic (GARCH) model, allowing us to capture time series

volatility over time. The GARCH model is typically denoted as GARCH(p, q), with parameters p and q, where p is the number of lag residuals errors and q is the number of lag variances. The ARIMA(p, d, q)-GARCH(p, q) model can be expressed as:

$$y_t = a + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

$$\varepsilon_t = \sqrt{\sigma_t} z_t \qquad \sigma^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 \tag{2}$$

where α_i, with $i = 1, \ldots, p$, and β_i, with $i = 1, \ldots, q$ are the parameters of the GARCH part of the model, and ε_t follows a Student's t distribution.

3.1.3 Google trends (`Google`)

Since the `Google` marginal in all locations includes several values equal to zero, we fitted a zero adjusted gamma distribution (ZAGA) using time as explanatory variable (see Rigby and Stasinopoulos (2005)). This distribution is a mixture of a discrete value 0 with probability ν, and a gamma distribution on the positive real line $(0, \infty)$ with probability $(1 - \nu)$. The probability density function (pdf) of the ZAGA model is given by

$$f_X(x|\mu, \sigma, \nu) = \begin{cases} \nu & \text{if } x = 0 \\ (1-\nu) f_{GA}(x|\mu, \sigma) & \text{if } x > 0 \end{cases} \tag{3}$$

for $0 \leq x < \infty$, $0 < \nu < 1$, where $\mu > 0$ is the scale parameter, $\sigma > 0$ is the shape parameter and $f_{GA}(x|\mu, \sigma)$ is the gamma pdf. We assumed that the parameter μ of the ZAGA model is related to time, as explanatory variable, through an appropriate link function, with coefficient β (for more details, see Rigby et al. (2019)).

3.1.4 Total number of tweets (`Total_tweets`)

The best fitting model for the marginal `Total_tweets` is the zero adjusted inverse Gaussian distribution (ZAIG), which is similar to the ZAGA model discussed in Section 3.1.3. The pdf of the ZAIG model is

$$f_X(x|\mu, \sigma, \nu) = \begin{cases} \nu & \text{if } x = 0 \\ (1-\nu) f_{IG}(x|\mu, \sigma) & \text{if } x > 0 \end{cases} \tag{4}$$

for $0 \leq x < \infty$, $0 < \nu < 1$, where $\mu > 0$ is the location parameter, $\sigma > 0$ is the scale parameter and $f_{IG}(x|\mu, \sigma)$ is the inverse Gaussian pdf. Similarly to the ZAGA

model, for the ZAIG model we assumed that the parameter μ is related to time, as explanatory variable, through an appropriate link function, with coefficient β (see Rigby and Stasinopoulos (2005); Rigby et al. (2019)).

3.1.5 Bing sentiment score (`Bing`)

The best model for the `Bing` marginal for all three locations was the ARIMA-GARCH model with Student's t innovations, as illustrated in Eq.(2), fitted on the log-transformed data.

Since the residuals of the Dawlish data still showed some structure, they were fitted using a Generalized t distribution (GT), which depends on four parameters controlling location, scale and kurtosis (for more information, see Rigby and Stasinopoulos (2005); Rigby et al. (2019)).

3.1.6 Afinn sentiment score (`Afinn`)

The log-transformed `Afinn` marginal was fitted with an ARIMA-GARCH model with Student's t innovations (see Eq.(2)).

For Portsmouth, since the residuals still presented some structure, they were fitted using a skew exponential power type 2 distribution (SEP2), which depends on four parameters: the location, scale, skewness and kurtosis. For the implementation of the SEP2 distribution, we used time as explanatory variable for the location parameter.

For Dawlish, the residuals were fitted using a Normal-exponential-Student-t distribution (NET), considering time as explanatory variable. The NET distribution is symmetric and depends on four parameters controlling the location, scale and kurtosis (for more details on the SEP2 and NET distributions, see Rigby and Stasinopoulos (2005); Rigby et al. (2019)).

3.2 Vine Copula Model

A *vine copula* (or *vine*) represents the pattern of dependence of multivariate data via a cascade of bivariate copulas, allowing us to construct flexible high-dimensional copulas using only bivariate copulas as building blocks. For more details about vine copulas see Czado (2019).

In order to obtain a vine copula we proceed as follows. First we factorise the joint distribution $f(x_1, \ldots, x_d)$ of the random vector $\mathbf{X} = (X_1, \ldots, X_d)$ as a product of conditional densities

$$f(x_1, \ldots, x_d) = f_d(x_d) \cdot f_{d-1|d}(x_{d-1}|x_d) \cdot \ldots \cdot f_{1|2\ldots d}(x_1|x_2, \ldots x_d). \tag{5}$$

The factorisation in (5) is unique up to re-labelling of the variables and it can be expressed in terms of a product of bivariate copulas. In fact, by Sklar's theorem, the conditional density of $X_{d-1}|X_d$ can be easily written as

$$f_{d-1|d}(x_{d-1}|x_d) = c_{d-1,d}(F_{d-1}(x_{d-1}), F_d(x_d); \boldsymbol{\theta}_{d-1,d}) \cdot f_{d-1}(x_{d-1}), \tag{6}$$

where $c_{d-1,d}$ is a bivariate copula, with parameter vector $\boldsymbol{\theta}_{d-1,d}$. Through a straightforward generalisation of Eq.(6), each term in (5) can be decomposed into the appropriate bivariate copula times a conditional marginal density. More precisely, for a generic element X_j of the vector $\mathbf{X}$ we obtain

$$f_{X_j|\mathbf{v}}(x_j|\mathbf{v}) = c_{X_J,\nu_\ell;\mathbf{v}_{-\ell}}(F_{X_j|\mathbf{v}_{-\ell}}(x_j|\mathbf{v}_{-\ell}), F_{\nu_\ell|\mathbf{v}_{-\ell}}(\nu_\ell|\mathbf{v}_{-\ell}); \boldsymbol{\theta}_{X_J,\nu_\ell;\mathbf{v}_{-\ell}}) \cdot f_{X_j|\mathbf{v}_{-\ell}}(x_j|\mathbf{v}_{-\ell}), \tag{7}$$

where $\mathbf{v}$ is the conditioning vector, ν_ℓ is a generic component of $\mathbf{v}$, $\mathbf{v}_{-\ell}$ is the vector $\mathbf{v}$ without the component ν_ℓ, $F_{X_j|\mathbf{v}_{-\ell}}(\cdot|\cdot)$ is the conditional distribution of X_j given $\mathbf{v}_{-\ell}$ and $c_{X_J,\nu_\ell;\mathbf{v}_{-\ell}}(\cdot,\cdot)$ is the conditional bivariate copula density with parameter $\boldsymbol{\theta}_{X_J,\nu_\ell;\mathbf{v}_{-\ell}}$. The d-dimensional joint multivariate distribution function can hence be expressed as a product of bivariate copulas and marginal distributions by recursively plugging Eq.(7) in Eq.(5).

For example, let us consider a 6-dimensional distribution. Then, Eq.(5) translates to

$$f(x_1, \ldots, x_6) = f_6(x_6) \cdot f_{5|6}(x_5|x_6) \cdot f_{4|5,6}(x_4|x_5, x_6) \cdot \ldots \cdot f_{1|2\ldots6}(x_1|x_2, \ldots x_6). \tag{8}$$

The second factor $f_{5|6}(x_5|x_6)$ on the right-hand side of (8) can be easily decomposed into the bivariate copula $c_{5,6}(F_5(x_5), F_6(x_6))$ and marginal density $f_5(x_5)$:

$$f_{5|6}(x_5|x_6) = c_{5,6}(F_5(x_5), F_6(x_6); \boldsymbol{\theta}_{5,6}) \cdot f_5(x_5).$$

On the other hand, the third factor on the right-hand side of (8) can be decomposed using the (7) as

$$f_{4|5,6}(x_4|x_5, x_6) = c_{4,5;6}(F_{4|6}(x_4|x_6), F_{5|6}(x_5|x_6); \boldsymbol{\theta}_{4,5;6}) \cdot f_{4|6}(x_4|x_6).$$

Therefore, one of the possible decompositions of the joint density $f(x_1, \ldots, x_6)$ is given by the following expression, which includes the product of marginal densities and copulas, which are all bivariate:

$$\begin{aligned} f(x_1, \ldots, x_6) = \prod_{j=1}^{6} f_j(x_j) \cdot c_{1,2} \cdot c_{1,3} \cdot c_{3,4} \cdot c_{1,5} \cdot c_{5,6} \cdot c_{2,3;1} \cdot c_{1,4;3} \cdot c_{3,5;1} \cdot c_{1,6;5} \\ \cdot c_{2,4;1,3} \cdot c_{4,5;1,3} \cdot c_{3,6;1,5} \cdot c_{2,5;1,3,4} \cdot c_{4,6;1,3,5} \cdot c_{2,6;1,3,4,5}. \end{aligned} \tag{9}$$

Eq.(9) is called *pair copula construction*. Note that in the previous equation the notation has been simplified, setting $c_{a,b} = c_{a,b}(F_a(x_a), F_b(x_b); \boldsymbol{\theta}_{a,b})$.

Pair copula constructions can be represented through a graphical model called *regular vine* (R-vine). An R-vine $\mathcal{V}(d)$ on d variables is a nested set of trees (connected acyclic graphs) $T_1, \ldots, T_{d-1}$, where the variables are represented by nodes linked by edges, each associated with a certain bivariate copula in the corresponding pair copula construction. The edges of tree T_k are the nodes of tree T_{k+1}, $k = 1, \ldots, d-1$. In an R-vine, if two edges are tree T_k share a common node, they are represented in tree T_{k+1} by nodes joined by an edge. Figure 4 shows the

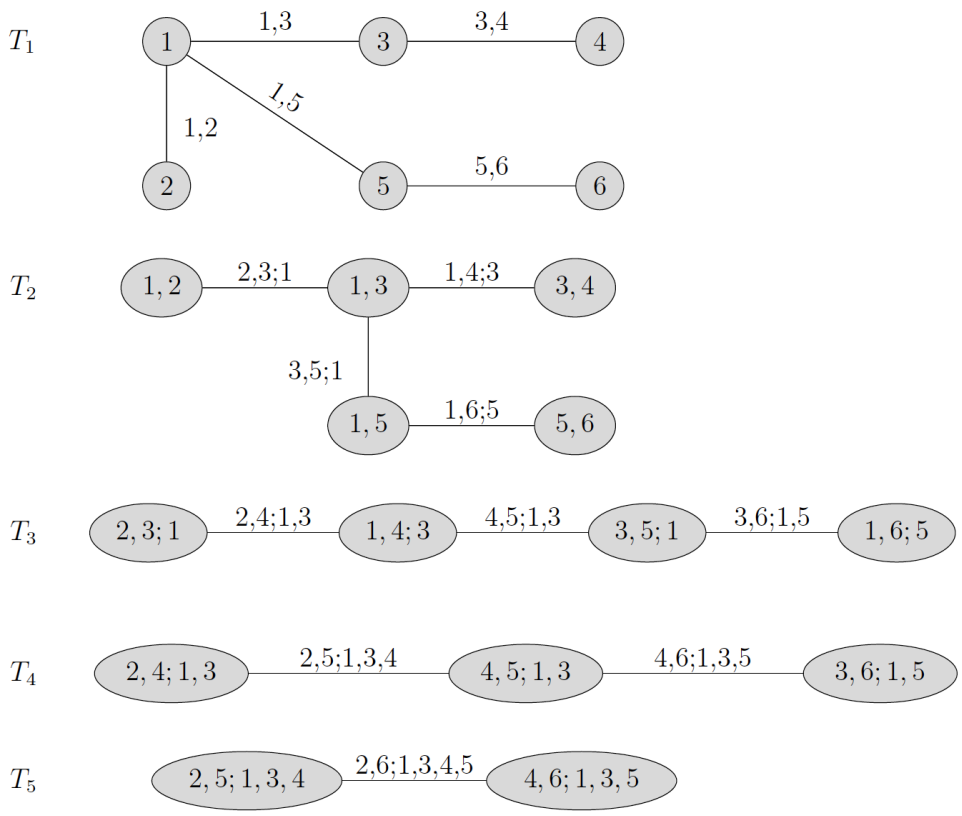

Figure 4: Six-dimensional R-vine graphical representation. *Source: Czado (2019)*

6-dimensional R-vine represented in Eq.(9). Each edge corresponds to a pair copula density (possibly belonging to different families) and the edge label corresponds to the subscript of the pair copula density, e.g. edge $2, 4; 1, 3$ corresponds to the copula $c_{2,4;1,3}$.

In order to estimate the vine, its structure as well as the copula parameters have to be specified. A sequential approach is generally adopted to select a suitable R-vine decomposition, specifying the first tree and then proceeding similarly for the following trees. For selecting the structure of each tree, we followed the approach suggested by Aas et al. (2009) and developed by Dissmann et al. (2013), using the maximal spanning tree algorithm. This algorithm defines a tree on all nodes (named spanning tree), which maximizes the sum of absolute pairwise dependencies, measured, for example, by Kendall's τ. This specification allows us to capture the strongest dependencies in the first tree and to obtain a more parsimonious model. Given the selected tree structure, a copula family for each pair of variables is identified using the Akaike Information Criterion (AIC), or the Bayesian Information Criterion (BIC). This choice is typically made amongst a large set of families, comprising elliptical copulas (Gaussian and Student's t) as well as Archimedean copulas (Clayton, Gumbel, Frank and Joe), their mixtures (BB1, BB6, BB7 and BB8) and their rotated versions, to cover a large range of possible dependence structures. For an overview of the different copula families, see Joe (1997) or Nelsen (2007). After specifying the vine structure and the copula families, the copula parameters $\boldsymbol{\theta}$ are then estimated using the maximum likelihood (MLE) method, as illustrated by Aas et al. (2009). The R-vine estimation procedure is repeated for all the trees, until the R-vine is completely specified.

4 Result Analysis and Discussion

We now present the results of the analysis of the remotely-sensed and online flood data for the three locations under consideration.

4.1 Twitter Wordclouds

First, we analysed the information gathered on Twitter, cleaning and stemming the tweets and producing wordclouds for each location.

Figure 5 displays the wordclouds of paired words obtained by pairing the the most common combinations of words appearing in the collected tweets. The top, middle and bottom panels show the wordclouds of Portsmouth, Plymouth and Dawlish tweets, respectively. It is interesting to see that the most frequent pairs of words refer to dates indicating storm and flood events (e.g. 28 October, 3 January), names of places affected by storms (e.g. Thorney Island, St Mary) and names of rivers (e.g. river Yealm, river Teign).

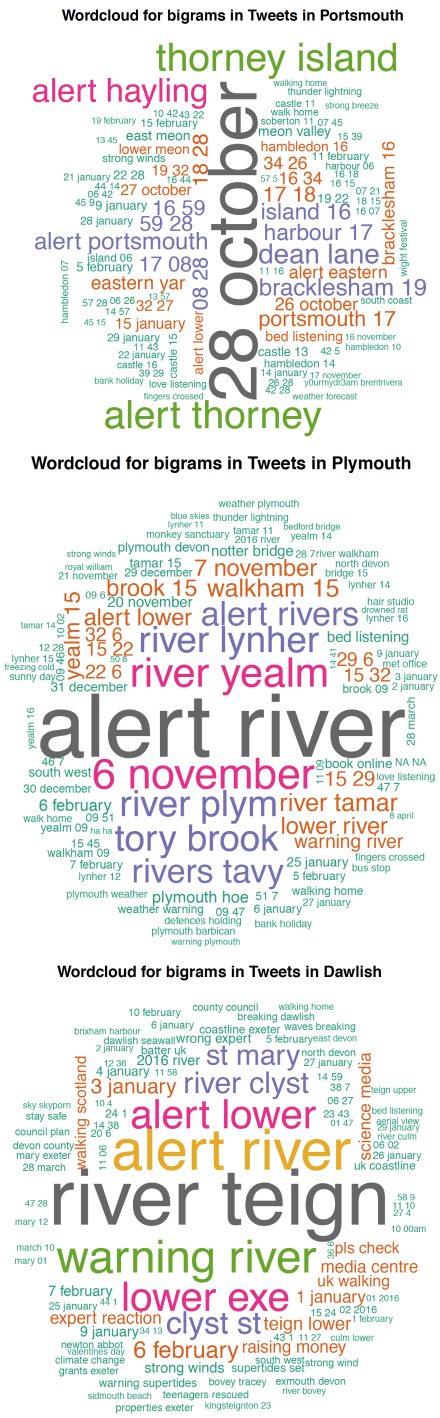

Figure 5: Wordcloud of paired words in tweets from Portsmouth (top panel), Plymouth (middle panel) and Dawlish (bottom panel).

4.2 Marginals Estimation

Table 2 lists the parameter estimates, obtained via the MLE method, of the best fitting models for the marginals, as described in Section 3.1, for Portsmouth (top panels), Plymouth (middle panels) and Dawlish (bottom panels). Standard errors are in brackets. [3]

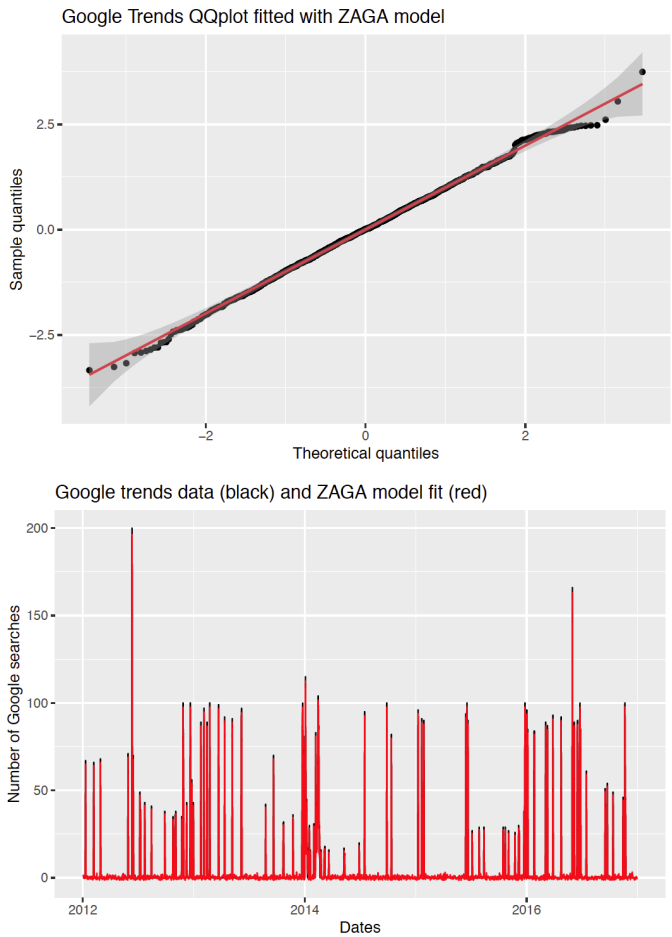

Figure 6: Plot illustrating the fit of the residuals for the `Google` marginal for Portsmouth. Top plot: QQ-plot comparing the Gaussian theoretical quantiles with sample quantiles. Bottom plot: observed time series (black line) and in-sample predictions obtained form the fitted ZAGA model (red line).

[3]Please, note that, due to lack of space, the Table does not include the estimates of the GT, SEP2 and NET models fitted to the residuals of the `Bing` and `Afinn` marginals.

Table 2: Parameter estimates of the marginals for each location: Portsmouth (top panels), Plymouth (middle panels) and Dawlish (bottom panels). Standard errors are in brackets.

Marginals											
Portsmouth											
`Hs`		`WL`		`Google`		`Total_tweets`		`Bing`		`Afinn`	
ARIMA(3,0,2)		ARIMA(1,0,1)-GARCH(1,1)		ZAGA		ZAIG		ARIMA(1,d,1)-GARCH(3,1)		ARIMA(1,d,0)-GARCH(3,1)	
a	-0.1146 (0.0730)	a	0.1373 (0.0066)	μ	2.2579 (1.5760)	μ	8.8258 (1.3940)	a	1.0000 (0.00009)	a	1.0000 (0.00009)
ϕ_1	0.9810 (0.2135)	ϕ_1	0.7679 (0.0221)	β	0.0001 (0.00009)	β	-0.0004 (0.00008)	ϕ_1	0.8707 (0.0172)	ϕ_1	0.3832 (0.0435)
ϕ_2	0.2891 (0.3474)	θ_1	-0.1174 (0.0351)	σ	-0.5203 (0.0581)	σ	-0.6741 (0.0197)	θ_1	-0.7335 (0.0351)	d	0.2836 (0.1092)
ϕ_3	-0.2836 (0.1382)	ω	0.0001 (0.00006)	ν	2.5440 (0.0901)	ν	-0.8606 (0.0512)	d	0.3382 (0.0363)	ω	0.0000 (0.0000)
θ_1	-0.3112 (0.2060)	α_1	0.0651 (0.0162)					ω	0.0000 (0.0000)	α_1	0.0167 (0.0038)
θ_2	-0.5775 (0.1979)	β_1	0.9214 (0.0196)					α_1	0.0167 (0.0037)	α_2	0.0167 (0.0097)
		σ	5.9142 (0.7878)					α_2	0.0167 (0.0086)	α_3	0.0167 (0.0075)
								α_3	0.0167 (0.0072)	β_1	0.9000 (0.0111)
								β_1	0.9000 (0.0096)	σ	3.9999 (0.2633)
								σ	4.0000 (0.2617)		
Plymouth											
ARIMA(1,0,2)		ARIMA(4,1,1)		ZAGA		ZAIG		ARIMA(1,d,2)-GARCH(2,1)		ARIMA(1,d,1)-GARCH(2,1)	
a	0.0048 (0.0473)	ϕ_1	0.7204 (0.0240)	μ	4.927 (1.403)	μ	4.826 (1.371)	a	1.0000 (0.00005)	a	1.0000 (0.00009)
ϕ_1	0.8366 (0.0248)	ϕ_2	0.0582 (0.0289)	β	-0.00005 (0.00008)	β	-0.0002 (0.00008)	ϕ_1	0.9779 (0.0007)	ϕ_1	0.3229 (0.0882)
θ_1	-0.1016 (0.0368)	ϕ_3	0.0035 (0.0289)	σ	-0.3348 (0.0395)	σ	-0.6058 (0.0219)	θ_1	-0.8854 (0.0115)	θ_1	-0.1301 (0.0968)
θ_2	-0.1598 (0.0337)	ϕ_4	0.0112 (0.0239)	ν	1.7263 (0.0653)	ν	-0.2788 (0.0473)	θ_2	-0.0553 (0.0115)	d	0.3111 (0.0683)
		θ_1	-0.9919 (0.0055)					d	0.3076 (0.0490)	ω	0.0000 (0.0000)
								ω	0.0000 (0.0000)	α_1	0.0250 (0.0062)
								α_1	0.0250 (0.0068)	α_2	0.0250 (0.0057)
								α_2	0.0250 (0.0051)	β_1	0.9000 (0.0137)
								β_1	0.9000 (0.0069)	σ	4.0000 (0.5849)
								σ	4.0000 (0.3481)		
Dawlish											
ARIMA(2,0,3)		ARIMA(1,0,1)-GARCH(1,1)		ZAGA		ZAIG		ARIMA(2,d,1)-GARCH(1,1)		ARIMA(1,d,1)-GARCH(1,1)	
a	-0.3644 (0.0763)	a	0.1131 (0.0078)	μ	0.3969 (2.2670)	μ	-0.2273 (1.910)	a	1.0000 (0.00008)	a	0.9999 (0.00009)
ϕ_1	1.5597 (0.1241)	ϕ_1	0.8011 (0.0180)	β	0.0002 (0.0001)	β	0.0015 (0.0001)	ϕ_1	-0.2924 (0.0315)	ϕ_1	0.3049 (0.0268)
ϕ_2	-0.5674 (0.1208)	θ_1	-0.0023 (0.0297)	σ	-0.0928 (0.0463)	σ	-0.6471 (0.0399)	ϕ_2	0.2188 (0.0374)	θ_1	0.0760 (0.0225)
θ_1	-0.9812 (0.1285)	ω	0.00008 (0.00004)	ν	2.1713 (0.0772)	ν	1.5725 (0.0620)	θ_1	0.7206 (0.0459)	d	0.3967 (0.0146)
θ_2	-0.0680 (0.0586)	α_1	0.0496 (0.0102)					d	0.3139 (0.0028)	ω	0.0000 (0.0000)
θ_3	0.0998 (0.0627)	β_1	0.9394 (0.0116)					ω	0.0000 (0.0000)	α_1	0.0500 (0.0024)
		σ	5.4261 (0.6958)					α_1	0.0500 (0.0023)	β_1	0.8999 (0.0019)
								β_1	0.8999 (0.0022)	σ	3.9997 (0.1728)
								σ	3.999 (0.1399)		

As an example, Figure 6 shows the fit of the residuals for the `Google` trends marginal for Portsmouth. The other plots for all marginals related to all three locations exhibit a similar behaviour. The top panel displays the QQ-plot comparing the Gaussian theoretical quantiles with the sample quantiles, while the bottom panel illustrates the observations (black line) and in-sample predictions obtained form the fitted ZAGA model (red line). The plots clearly show an excellent fit of the ZAGA model to the marginal, as demonstrated by the points in the QQ-plot aligning almost perfectly to the main diagonal and the in-sample predictions overlapping the observed data.

4.3 Vine Estimation

Once estimated the marginals, we derived the corresponding *u-data* from the residuals, as illustrated in Section 3.1. Then, we carried out fitting and model selection for the vine copula for each location using the R package `rvinecopulib` (Nagler and Vatter, 2021).

Figure 7 displays the first trees of the vine copulas estimated for Portsmouth (top panel), Plymouth (middle panel) and Dawlish (bottom panel). The nodes are denoted with blue dots, with the names of the margins reported in boldface[4]. On each edge, the plots show the name of the selected pair copula family and the estimated copula parameter expressed as Kendall's τ. In order to estimate the vines, we adopted the Kendall's τ criterion for tree selection, the AIC for the copula families selection and the MLE method for estimating the pair copula parameters. As it is clear from Figure 7, the vines for the three different locations exhibit a very similar structure, with the environmental variables `Hs` and `WL` playing a central role and linking to the social media variables. The sentiment scores `Bing` and `Afinn` are directly associated. Likewise, `Total_tweets` and `Google` are contiguouly related. The symmetric Gaussian copula, which is often employed in traditional multivariate modelling, was not identified as the best fitting copula for neither of the locations. On the contrary, the selected copula families include the Student's t copula, which is able to model strong tail dependence, Archimedean copulas such as the Clayton and Gumbel, that are able to capture asymmetric dependence, and mixture copulas such as the BB1 (Clayton-Gumbel) and BB8 (Joe-Frank), that can accommodate various dependence shapes. Most of the associations between the variables are positive. The strongest associations are between the `Bing` and `Afinn` sentiment scores and between the environmental variables `Hs` and `WL`. Also, `Hs` and `Total_tweets` are mildly associated.

[4]Please, note that `Total_tweets` is denoted with `Tw` in the plots.

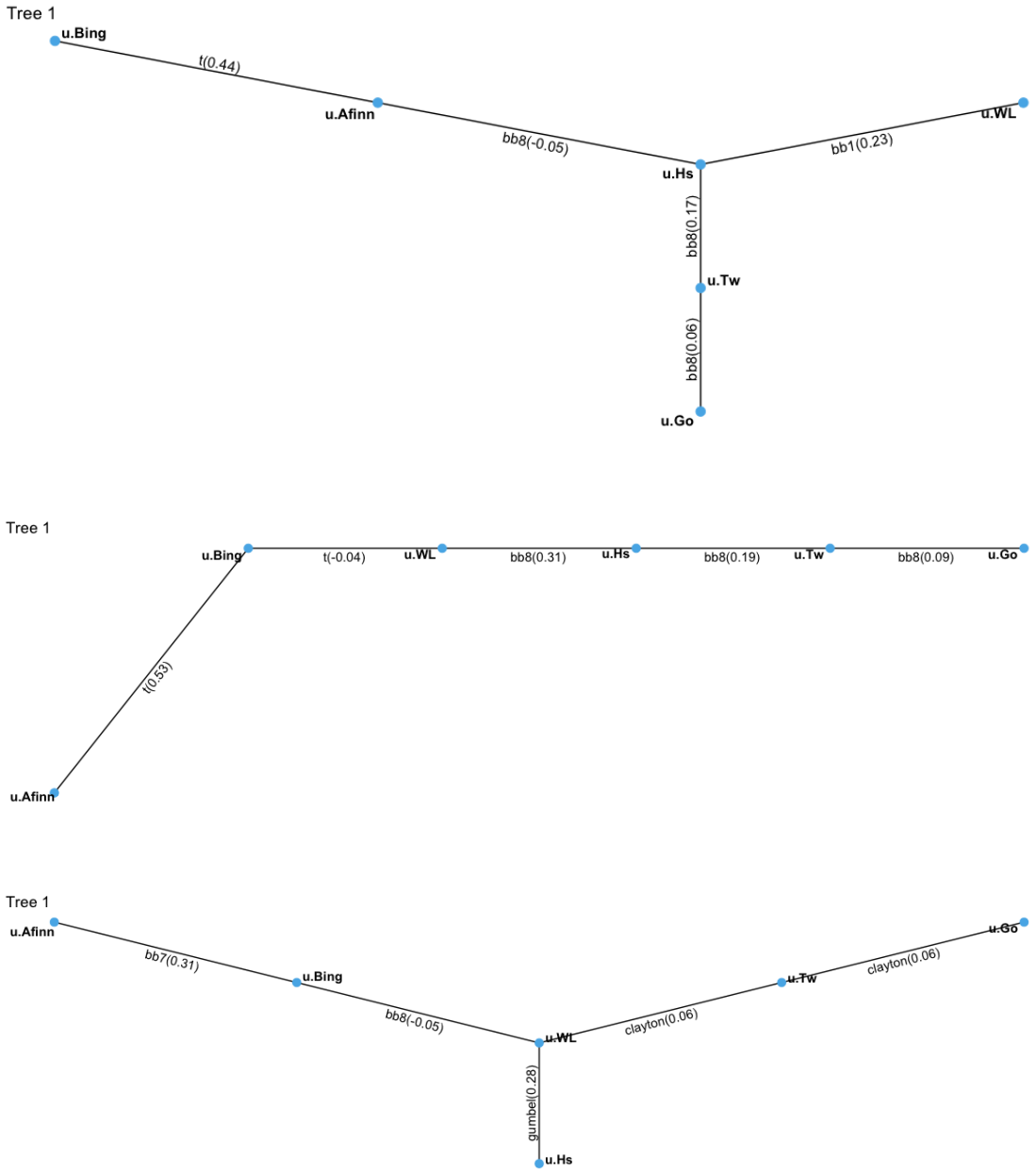

Figure 7: First trees of the vine copulas estimated for Portsmouth (top panel), Plymouth (middle panel) and Dawlish (bottom panel).

4.4 Out-of-sample predictions

In this Section we test the predictive power of the proposed vine methodology, which integrates environmental and social media variables, constructing out-of-sample predictions. We also compare the predictions obtained with our methodology with those yielded using two traditional approaches. The former is based on vines built exclusively using Gaussian pair copulas, which are the most common in applications, but are restricted to dependence symmetry and absence of tail dependence. The latter approach assumes independence among the six time series under consideration and therefore calculates predictions ignoring any association between environmental and online information.

Out-of-sample predictions based on the proposed model were constructed considering the vine copula estimated as illustrated in Section 4.3 until the 15th February 2016 and using it to predict the period between the 16th February 2016 and the 31st December 2016. Let $\mathbf{X} = \{\mathbf{X}_t; t = 1, .., T\}$ be the 6-dimensional time series of environmental and social media data. Our aim is to forecast $\mathbf{X}_{T+1}$ based on the information available at time T. In order to do that, we adopted the forecasting method described by Simard and Rémillard (2015). We first extracted the residuals form the marginals, as explained in Section 3.1, and obtained the *u-data*. Next, we simulated M realizations from the vine copula. Hence, we calculated the predicted values for each simulation, using the inverse cdf and the relevant fitted marginal models. Then, we calculated the average prediction for all simulations $\hat{\mathbf{X}}_{T+1}$ and use it to forecast $\mathbf{X}_{T+1}$. The prediction interval of level $(1-\alpha) \in (0,1)$ for $\mathbf{X}_{T+1}$ was calculated by taking the estimated quantiles of order $\alpha/2$ and $1-\alpha/2$ amongst the simulated data. We denote by $\hat{\mathbf{X}}^l_{T+1}$ and $\hat{\mathbf{X}}^u_{T+1}$ the lower and upper values of the prediction intervals.

In order to assess the accuracy of our predictions, we made use of the mean squared error (MSE) to evaluate point forecasts and of the mean interval score (MIS), proposed by Gneiting and Raftery (2007), to evaluate the accuracy of the prediction intervals. The MSE for each variable $j = 1, \ldots, d$ was calculated as follows

$$\text{MSE}_j = \frac{1}{S} \sum_{t=T+1}^{T+S} (x_{t,j} - \hat{x}_{t,j})^2$$

where $x_{t,j}$ is the observed value for each variable at each time point t, $\hat{x}_{t,j}$ is the corresponding predicted value, $T+1$ denotes the 16th February 2016, while $T+S$ indicates the 31st December 2016. The 95% MIS for each variable, at level $\alpha = 0.05$,

was computed as

$$\mathrm{MIS}_j = \frac{1}{S}\sum_{t=T+1}^{T+S}\left[(\hat{x}^u_{t,j} - \hat{x}^l_{t,j}) + \frac{2}{\alpha}(\hat{x}^l_{t,j} - x_{t,j})\mathbb{1}(x_{t,j} < \hat{x}^l_{t,j}) + \frac{2}{\alpha}(x_{t,j} - \hat{x}^u_{t,j})\mathbb{1}(x_{t,j} > \hat{x}^u_{t,j})\right]$$

where $\hat{x}^l_{t,j}$ and $\hat{x}^u_{t,j}$ denote, respectively, the lower and upper limits of the prediction intervals for each variable at each time point, and $\mathbb{1}(\cdot)$ is the indicator function.

Tables 3 and 4 list the MSE and MIS values calculated for Portsmouth, Plymouth and Dawlish, in the top, middle and bottom panel, respectively, for each variable. The second columns show the vine copula results, the third columns show the results assuming all Gaussian pair-copulas, and the fourth columns show the results assuming independence among variables. The MSEs and MISs of the best performing approaches for each variable are highlighted in boldface. From Table 3, we notice that, in terms of MSE, the vine copula approach outperforms the other two approaches in the majority of the cases. In terms of MIS, Table 4 shows that the Gaussian and vine copula approaches always perform better than the traditional independence approach, except for two marginals related to Dawlish. Comparing the three different locations, in Plymouth the vine copula exceeds the performance of the other two approaches for most of the variables, whereas the independence approach is never selected. In Portsmouth the Gaussian vine method achieves generally the best results, with the independence approach only selected in two cases by the MSE indicator. In Dawlish, the vine and Gaussian copula methods are preferred for several variables, although the independence approach is selected in a few cases. This might be due to the lack of social media information for Dawlish, compared to the other two locations, as shown in Figure 3, making it difficult to define associations between online and environmental data and to leverage data integration for predicting purposes.

The variables `Hs` and `WL` are generally better predicted by the vine method, demonstrating that the use of online-generated information is able to improve the forecasts of environmental variables and that, ultimately, social media data integration allows us to obtain more accurate predictions of inundations and flood events.

The prediction of online-generated information also benefits from data integration. `Google` trends are more accurately forecasted by the vine copula method, or by the Gaussian approach in the Portsmouth case, rather than by the independent approach. The prediction of `Tot_tweets` achieves better results with the vine copula method for Plymouth data and with the Gaussian method for Portsmouth data, while the independence approach is selected only for Dawlish data, due to the lack of information for this location, as explained above.

Table 3: MSEs calculated for Portsmouth (top panel), Plymouth (middle panel) and Dawlish (bottom panel) for each variable. The figures show the vine copula results (second column), the results assuming all Gaussian pair-copulas (third column), and assuming independence among variables (fourth column). The MSEs of the best performing approaches for each variable are in boldface.

MSE Portsmouth			
Variable	Vine Copula	Gaussian	Independent
`Hs`	**1.2552**	1.2629	1.2703
`WL`	0.0299	0.0298	**0.0275**
`Google`	404.4147	**403.9977**	404.4304
`Total_Tweets`	6.7829	**6.6994**	6.7351
`Bing`	2.6624×10^{-11}	2.6634×10^{-11}	$\mathbf{2.6572 \times 10^{-11}}$
`Afinn`	1.3767×10^{-10}	$\mathbf{1.3745 \times 10^{-10}}$	1.3823×10^{-10}

MSE Plymouth			
Variable	Vine Copula	Gaussian	Independent
`Hs`	**1.3888**	1.4029	1.3901
`WL`	**0.02617**	0.02725	0.0272
`Google`	**2873.466**	2875.053	2874.761
`Total_Tweets`	**14.1569**	14.1698	14.2388
`Bing`	2.6829×10^{-11}	$\mathbf{2.6282 \times 10^{-11}}$	2.6834×10^{-11}
`Afinn`	$\mathbf{1.2028 \times 10^{-10}}$	1.2035×10^{-10}	1.2103×10^{-10}

MSE Dawlish			
Variable	Vine Copula	Gaussian	Independent
`Hs`	1.5713	1.5714	**1.5499**
`WL`	0.0252	0.0258	**0.0237**
`Google`	**4612.738**	4613.572	4612.772
`Total_Tweets`	610.3111	610.042	**609.9969**
`Bing`	5.6264×10^{-9}	$\mathbf{5.6124 \times 10^{-9}}$	5.7304×10^{-9}
`Afinn`	$\mathbf{6.1208 \times 10^{-9}}$	6.1670×10^{-9}	6.1873×10^{-9}

Table 4: MISs calculated for Portsmouth (top panel), Plymouth (middle panel) and Dawlish (bottom panel) for each variable. The figures show the vine copula results (second column), the results assuming all Gaussian pair-copulas (third column), and assuming independence among variables (fourth column). The MISs of the best performing approaches for each variable are in boldface.

MIS Portsmouth			
Variable	Vine Copula	Gaussian	Independent
`Hs`	**0.2437**	0.2441	0.2451
`WL`	0.04131	**0.0391**	0.0394
`Google`	6.1169	**6.1151**	6.1179
`Total_Tweets`	0.6356	**0.6316**	0.6366
`Bing`	1.2039×10^{-6}	$\mathbf{1.1982 \times 10^{-6}}$	1.2021×10^{-6}
`Afinn`	$\mathbf{3.1644 \times 10^{-6}}$	3.1668×10^{-6}	3.175×10^{-6}

MIS Plymouth			
Variable	Vine Copula	Gaussian	Independent
`Hs`	**0.2476**	0.2509	0.2491
`WL`	0.0375	**0.0374**	0.0388
`Google`	**10.8759**	10.879	10.8789
`Total_Tweets`	**0.7833**	0.7845	0.7849
`Bing`	1.2175×10^{-6}	$\mathbf{1.2043 \times 10^{-6}}$	1.2178×10^{-6}
`Afinn`	3.0704×10^{-6}	$\mathbf{3.0693 \times 10^{-6}}$	3.0799×10^{-6}

MIS Dawlish			
Variable	Vine Copula	Gaussian	Independent
`Hs`	0.2628	**0.2588**	0.2603
`WL`	0.0361	0.0357	**0.0350**
`Google`	**13.5782**	13.5794	13.5782
`Total_Tweets`	7.0913	7.0889	**7.0887**
`Bing`	1.7284×10^{-5}	$\mathbf{1{,}7184 \times 10^{-5}}$	1.7472×10^{-5}
`Afinn`	$\mathbf{1.8301 \times 10^{-5}}$	1.8385×10^{-5}	1.8396×10^{-5}

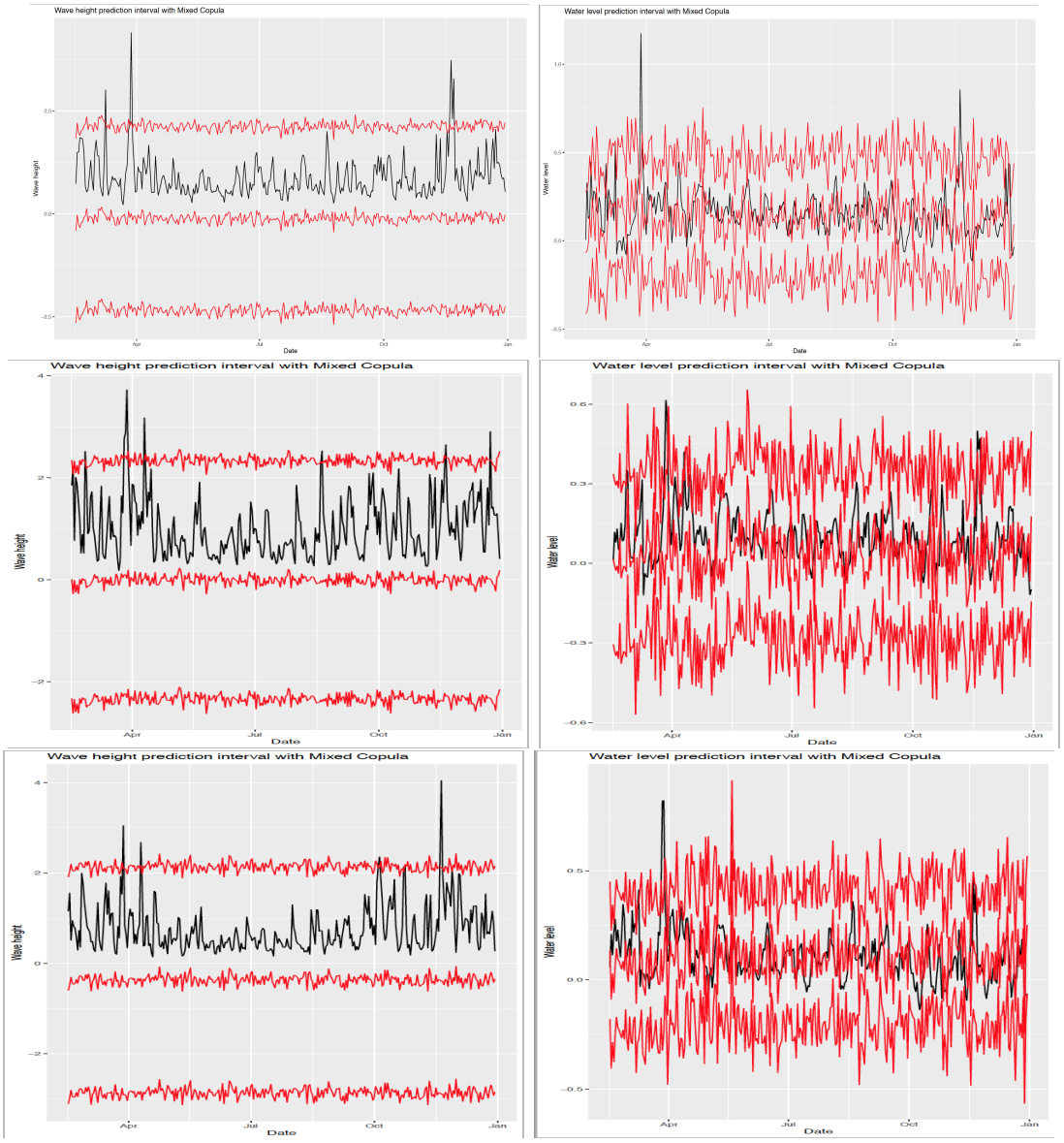

Figure 8: Line plots showing forecasts and prediction intervals for `Hs` (left panels) and `WL` (right panel) obtained with the vine copula methodology for the period between the 16th February 2016 and the 31st December 2016, for Portsmouth (top panel), Plymouth (middle panel) and Dawlish (bottom panel). Observed values are in black, predicted values are the inner red lines and 95% prediction intervals are the outer red lines.

Comparing the sentiment scores, we notice that the vine copula approach is generally preferred with `Afinn`, while the Gaussian method is typically selected with `Bing`. This is probably due to the fact that the `Afinn` lexicon is more sophisticated than `Bing`, since it scores words into several positive and negative categories, and hence provides more information.

Figure 8 shows the forecasts and prediction intervals for the wave height `Hs` and water level `WL` (on the left and right panel, respectively), obtained with the vine copula methodology for the period between the 16th February 2016 and the 31st December 2016. The top panels depict the Portsmouth plots, the middle panels depict the Plymouth plots and the bottom panels depict the Dawlish plots. The black lines denote the observed values, the inner red lines denote the predicted values and the outer red lines denote the 95% prediction intervals. We notice that the intervals predicted with the vine copula method capture most of the dynamic of the environmental variables, indicating that the proposed methodology is able to leverage social media information to provide accurate predictions of flood-related data.

5 Concluding Remarks

In this paper, we proposed a new methodology aiming at obtaining more accurate forecasts of variables measuring inundations and floods events. The proposed methodology is based on the integration of environmental variables collected via remote sensing, with online generated social media information. We collected data at three different locations in the South coast of the UK, which were affected by severe storm events in several occasions in the past few years. Together with wave height and water level information, we also gathered Google Trends searches and Twitter micro-blogging messages involving keywords related to floods and storms. From the tweets, we considered the volume as well as the sentiment scores, to investigate the feelings of people towards inundation events. Our methodology is based on vine copulas, which are able to model the dependence structure between the marginals, and thus to take advantage of the association between social media and environmental variables. The methodology is articulated in two steps. In the first step, the variables are modelled through time series analysis, in order to remove the effects of time dynamics from the margins. We selected the best fitting time series model for each variable, which were generally different for specific marginals, due to the distinct nature and behaviour of various variables. Then, in the second step, residuals were extracted and transformed to construct the vine copula model. The flexibility of vines allowed us to build a graphical structure based on different bivariate pair-copulas copulas, embedding the

dependence structure between the marginals. The bivariate pair-copulas included asymmetric and tail-dependent families, showing the need for more flexible than traditional approaches to model the dependencies between variables. We tested our approach calculating out-of-sample predictions and comparing the vine copula method with two traditional approaches: the first based on a vine constructed with all Gaussian copulas, and the second based on independence between variables. The results show that the vine copula method outperforms the other two approaches in most cases, demonstrating that our methodology is able to leverage social media information to obtain more accurate predictions of floods and inundations. In some cases, the Gaussian vine copula method is selected, showing that the vine data integration approach is still achieving the best performance, although some variables are less affected by asymmetries and tail dependence. Since social media information for Dawlish were lacking, they provided a more limited contribution to the prediction of the environmental variables for this location.

Further investigations involving other locations and including additional social media information will be the object of future work. Another extension will involve Bayesian inference, which would allow us to incorporate other information, such as experts' opinion, in the model.

Acknowledgements

This work was supported by the European Regional Development Fund project *Environmental Futures & Big Data Impact Lab*, funded by the European Structural and Investment Funds, grant number 16R16P01302 .

References

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics 44*(2), 182–198.

Alam, F., F. Ofli, and M. Imran (2018). Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 12.

Arthur, R., C. A. Boulton, H. Shotton, and H. T. Williams (2018). Social sensing of floods in the uk. *PloS one 13*(1), e0189327.

Balogun, A., S. Quan, B. Pradhan, U. Dano, and S. Yekeen (2020). An improved flood susceptibility model for assessing the correlation of flood hazard and property prices using geospatial technology and fuzzy-anp. *Journal of Environmental Informatics*.

Brouwer, T., D. Eilander, A. v. Loenen, M. J. Booij, K. M. Wijnberg, J. S. Verkade, and J. Wagemaker (2017). Probabilistic flood extent estimates from social media flood observations. *Natural Hazards and Earth System Sciences 17*(5), 735–747.

Czado, C. (2019). Analyzing dependent data with vine copulas. *Lecture Notes in Statistics, Springer*.

De Albuquerque, J. P., B. Herfort, A. Brenning, and A. Zipf (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International journal of geographical information science 29*(4), 667–689.

Dissmann, J., E. C. Brechmann, C. Czado, and D. Kurowicka (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis 59*, 52–69.

Feng, Y., P. Shi, S. Qu, S. Mou, C. Chen, and F. Dong (2020). Nonstationary flood coincidence risk analysis using time-varying copula functions. *Scientific reports 10*(1), 1–12.

Field, C. B., V. Barros, T. F. Stocker, and Q. Dahe (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association 102*(477), 359–378.

Grego, J. M., P. A. Yates, and K. Mai (2015). Standard error estimation for mixed flood distributions with historic maxima. *Environmetrics 26*(3), 229–242.

Herfort, B., J. P. de Albuquerque, S.-J. Schelhorn, and A. Zipf (2014). Exploring the geographical relations between social media and flood phenomena to improve situational awareness. In *Connecting a digital Europe through location and place*, pp. 55–71. Springer.

Hu, M. and B. Liu (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177.

Huang, K., L. Dai, M. Yao, Y. Fan, and X. Kong (2017). Modelling dependence between traffic noise and traffic flow through an entropy-copula method. *Journal of Environmental Informatics 29*(2).

Hyndman, R. J. and G. Athanasopoulos (2018). *Forecasting: principles and practice*. OTexts.

Jane, R., L. Dalla Valle, D. Simmonds, and A. Raby (2016). A copula-based approach for the estimation of wave height records through spatial correlation. *Coastal Engineering 117*, 1–18.

Jane, R. A., D. J. Simmonds, B. P. Gouldby, J. D. Simm, L. Dalla Valle, and A. C. Raby (2018). Exploring the potential for multivariate fragility representations to alter flood risk estimates. *Risk Analysis 38*(9), 1847–1870.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Joe, H. and J. J. Xu (1996). The estimation method of inference functions for margins for multivariate models. Technical Report 166, Department of Statistics, University of British Columbia.

Keef, C., J. A. Tawn, and R. Lamb (2013). Estimating the probability of widespread flood events. *Environmetrics 24*(1), 13–21.

Latif, S. and F. Mustafa (2020). Parametric vine copula construction for flood analysis for kelantan river basin in malaysia. *Civil Engineering Journal 6*(8), 1470–1491.

Li, Z., C. Wang, C. T. Emrich, and D. Guo (2018). A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 south carolina floods. *Cartography and Geographic Information Science 45*(2), 97–110.

Mason, D. C., I. J. Davenport, J. C. Neal, G. J.-P. Schumann, and P. D. Bates (2012). Near real-time flood detection in urban and rural areas using high-resolution synthetic aperture radar images. *IEEE transactions on Geoscience and Remote Sensing 50*(8), 3041–3052.

Massicotte, P. and D. Eddelbuettel (2021). *gtrendsR: Perform and Display Google Trends Queries*. R package version 1.4.8.

Moishin, M., R. C. Deo, R. Prasad, N. Raj, and S. Abdulla (2020). Development of flood monitoring index for daily flood risk evaluation: case studies in fiji. *Stochastic Environmental Research and Risk Assessment*, 1–16.

Muller, C., L. Chapman, S. Johnston, C. Kidd, S. Illingworth, G. Foody, A. Overeem, and R. Leigh (2015). Crowdsourcing for climate and atmospheric sciences: Current status and future potential. *International Journal of Climatology 35*(11), 3185–3203.

Nagler, T. and T. Vatter (2021). *rvinecopulib: High Performance Algorithms for Vine Copula Modeling*. R package version 0.5.5.1.1.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(3), 507–554.

Rigby, R. A., M. D. Stasinopoulos, G. Z. Heller, and F. De Bastiani (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. CRC press.

Rosser, J. F., D. Leibovici, and M. Jackson (2017). Rapid flood inundation mapping using social media, remote sensing and topographic data. *Natural Hazards 87*(1), 103–120.

Saravanou, A., G. Valkanas, D. Gunopulos, and G. Andrienko (2015). Twitter floods when it rains: a case study of the uk floods in early 2014. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1233–1238.

Schumann, G., P. D. Bates, M. S. Horritt, P. Matgen, and F. Pappenberger (2009). Progress in integration of remote sensing–derived flood extent and stage data and hydraulic models. *Reviews of Geophysics 47*(4).

Silge, J. and D. Robinson (2016). tidytext: Text mining and analysis using tidy data principles in R. *Journal of Statistical Software 1*(3).

Simard, C. and B. Rémillard (2015). Forecasting time series with multivariate copulas. *Dependence modeling 3*(1).

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris 8*, 229–231.

Smith, L., Q. Liang, P. James, and W. Lin (2017). Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *Journal of Flood Risk Management 10*(3), 370–380.

Spielhofer, T., R. Greenlaw, D. Markham, and A. Hahne (2016). Data mining twitter during the uk floods: Investigating the potential use of social media in emergency management. In *2016 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pp. 1–6. IEEE.

Spruce, M. D., R. Arthur, J. Robbins, and H. T. Williams (2021). Social sensing of high-impact rainfall events worldwide: A benchmark comparison against manually curated impact observations. *Natural Hazards and Earth System Sciences Discussions*, 1–31.

Talukdar, S., B. Ghose, R. Salam, S. Mahato, Q. B. Pham, N. T. T. Linh, R. Costache, M. Avand, et al. (2020). Flood susceptibility modeling in teesta river basin, bangladesh using novel ensembles of bagging algorithms. *Stochastic Environmental Research and Risk Assessment 34*(12), 2277–2300.

Tosunoglu, F., F. Gürbüz, and M. N. İspirli (2020). Multivariate modeling of flood characteristics using vine copulas. *Environmental Earth Sciences 79*(19), 1–21.

UN (2015). The human cost of weather related disasters 1995–2015, United Nations, Geneva, Switzerland, 30 pp.

Wang, X. and C. Du (2003). An internet based flood warning system. *Journal of Environmental Informatics 2*(1), 48–56.