ISEIS

Journal of Environmental Informatics 39(2) 97-110 (2022)

Journal of Environmental Informatics

www.iseis.org/jei

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

L. Ansell[1], and L. Dalla Valle[1 *]

[1] *School of Engineering, Computing and Mathematics, Plymouth University, Plymouth PL4 8AA, UK*

Received 05 April 2021; revised 20 May 2021; accepted 01 October 2021; published online 05 January 2022

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 on the South coast of the United Kingdom (UK). The results show that our approach, based on the integration of social media data, outperforms traditional methods in terms of evaluation and prediction of flood events.

Keywords: climate change, dependence modelling, floods, natural hazards, social media sentiment analysis, time series modelling, vine copulas

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 residing in locations experiencing inundations 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, in order to estimate and predict inundations, statistical and machine learning models are employed, typically using information 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 (2004). 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. Model validation results showed that 59.42 and 36.23% of past flood events fall within the very high and high susceptible locations of the susceptibility map respectively. 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. Results showed that an area of more than 800 km^2 was predicted as a very high flood susceptibility zone by all algorithms.

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 dis-

* Corresponding author. Tel.: 01752 586319.
E-mail address: luciana.dallavalle@plymouth.ac.uk (L. Dalla Valle).

ISSN: 1726-2135 print/1684-8799 online
 doi:10.3808/jei.202200471

semination 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 have 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. The authors evaluated their findings against ground truth data, obtained from external independent sources, and were able to identify flood-stricken areas. 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. Results show that the authors' approach could provide a consistent and comparable estimation of the flood situation in near real time. 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. The authors showed that social sensing successfully identified most high-impact rainfall events present in the manually curated database, with an overall accuracy of 95%.

However, the majority of contributions in the literature analysing online generated data focus 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 environmental information via sophisticated statistical modelling based on vine copulas, to enhance the assessment and prediction of flood phenomena compared to traditional approaches.

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 3-dimensional loading parameter space, estimating the overall failure probability. Copulas have also been applied in a flood risk context to model the dependence between multiple co-occurring drivers by Ward et al. (2018), among others. Couasnon et al. (2018) use Gaussian pair-copulas in a Bayesian Network to derive boundary conditions that account for riverine and coastal interactions for a catchment in southeast Texas. Feng et al. (2020) employed 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 modelling complex dependence structures between the variables. Vine copulas were adopted, for example, by Latif and Mustafa (2020) to model trivariate flood characteristics for the Kelantan River basin in Malaysia. Tosunoglu et al. (2020) applied vine copulas in hydrology for multivariate modelling of peak, volume and duration of floods in the Euphrates River Basin, Turkey. Vine copulas were applied to model compound events by Bevacqua et al. (2017) and by Santos et al. (2021). The former authors adopted this approach to quantify the risk in present-day and future climate, and to measure uncertainty estimates around such risk. The latter authors used vines to assess compound flooding from storm surge and multiple riverine discharges in Sabine Lake, Texas.

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 proposed methodology involves the use of subjective information, more specifically the feelings of people expressed through social media and quantified by sentiment scores, not merely as stand-alone data sources, used in isolation to predict inundations, but combined with information on the occurrence and magnitude of flood events. The vine copula approach allows us to exploit the associations between all the considered data sources, environmental as well as on-line, which all contribute to calculate flood forecasts.

The methodology articulates in the following steps, that will be illustrated in detail in the following sections:

- fit each variable (environmental as well as on-line information) with a suitable time series model, to remove the temporal effects from the data;
- construct a vine copula model, which accounts for the dependencies between all variables and exploits the associations between environmental and social media information;
- calculate predictions of the flood variables based on the vine copula model.

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, which do not take into account associations between environmental and on-line information, to estimate and predict the occurrence and the magnitude of 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 on the South coast of the UK, which were severely affected by storm events in recent years: Portsmouth, Plymouth and Dawlish. The inundation episodes of the last few years had a substantial socioeconomic 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. The seawall was breached, a temporary line of shipping containers forming a breakwater was constructed, however huge waves damaged it and punched a new hole in the sea wall. Later, a replacement seawall was installed and railway operations recommenced on April 4, 2014. The waves on the night of the 4th February were relatively modest. The breach was more likely a result of a combination of factors including coincidental arrival of swell waves and the highest locally generated wind waves, large storm surge arriving a few days after a spring tide and the sequence of storm events hitting the South UK coast that winter before the breach lowering beach level (Sibley et al., 2015).

In order to 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.

For each one of the three locations, we obtained 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. We therefore constructed a dataset of time series, all of the same length. More precisely, we downloaded wave height (m) and water level (tidal residual, m) data from the UK Environment Agency flood-monitoring API. 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 gtrendsR package from the R software (R Core Team, 2020; Massicotte and Eddelbuettel, 2021). 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 score. In order to take into account of the different population sizes living in the three areas, we scaled the Bing and Afinn sentiment scores by the relevant number of residents.

Table 1. Major Winter Storm Events in the UK Between 2012 and 2018

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						

Note: the storm naming system was introduced in 2015.

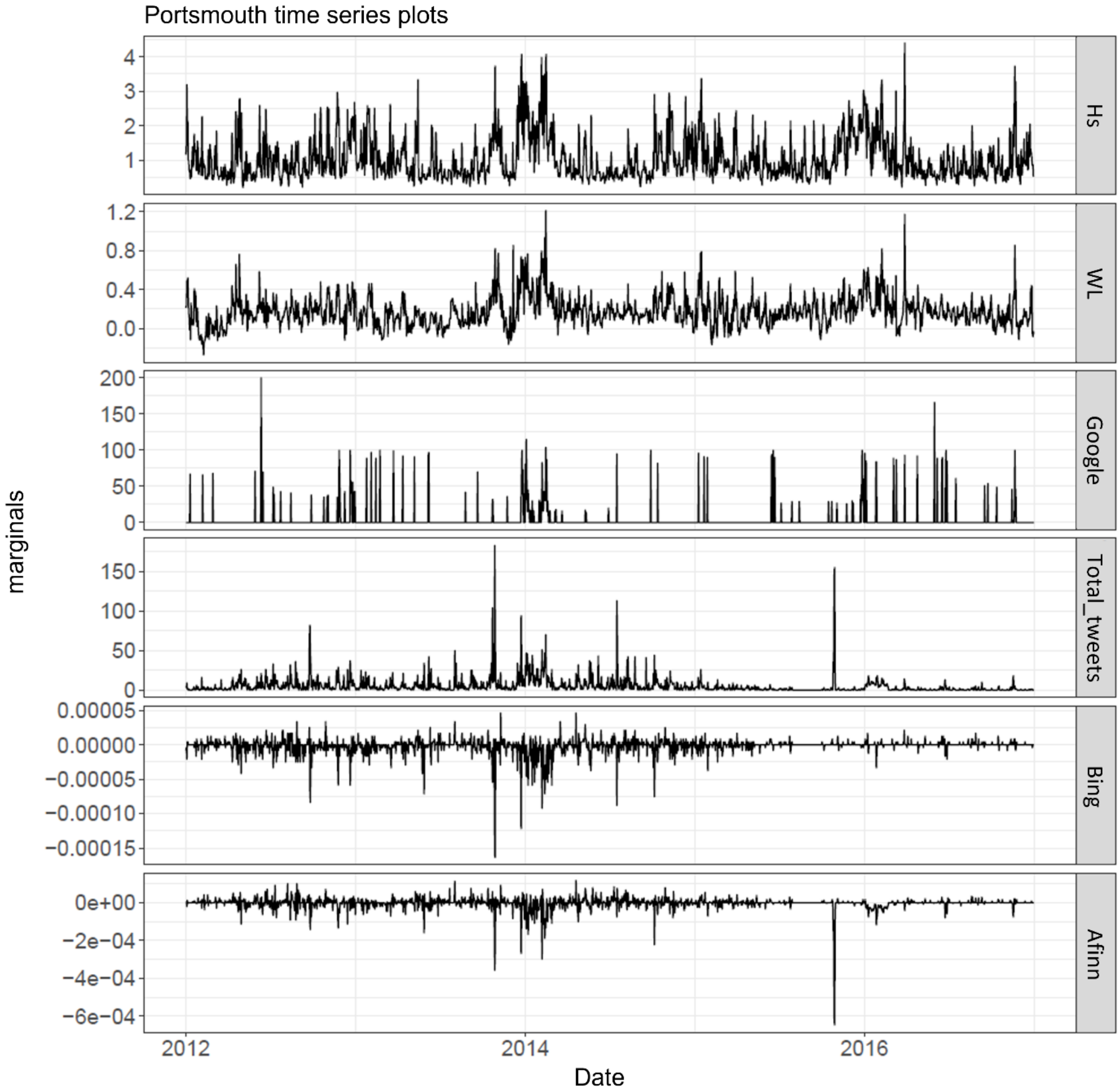

Figure 1. Trace plots of Portsmouth data.

Figure 1 and Figures S1 and S2 in the supplementary materials show the trace plots of the data collected for Portsmouth, Plymouth and Dawlish, respectively. The plots are produced using a daily scale. 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 in February 2014 are reflected in high spikes in the time series plots, especially for Dawlish in Figure S2. 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 suggests 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 copula theory. It states that, given a vector of random variables $\mathbf{X} = (X_1, ..., X_d)$, with d-dimensional joint cumulative distribution function $F(x_1, ..., x_d)$ and marginal cumulative distributions (cdf) $F_j(x_j)$, with $j = 1, ..., d$, a d-dimensional copula C exists, such that:

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d); \vartheta) \tag{1}$$

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

$$f(x_1,\ldots,x_d) = c(F_1(x_1),\ldots,F_d(x_d);\vartheta)\cdot f_1(x_1)\cdot\ldots\cdot f_d(x_d) \quad (2)$$

where c denotes the d-variate copula density. The copula allows us to determine the joint multivariate distribution and 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 $j = 1, ..., d$, 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 $t = 1, ..., T$ takes the following form:

$$y_t = a + \sum_{i=1}^{p}\varphi_i y_{t-1} + \sum_{i=1}^{q}\theta_i\varepsilon_{t-1} + \varepsilon_t \quad (3)$$

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 $i = 1, ..., p$, are the autoregressive parameters, θ_i, with $i = 1, ..., q$, 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 Equation (3). However, for Portsmouth and Dawlish, the ARIMA-GARCH model with Student's t innovations appeared to exhibit 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:

$$y_t = a + \sum_{i=1}^{p}\varphi_i y_{t-1} + \sum_{i=1}^{q}\theta_i\varepsilon_{t-i} + \varepsilon_t,$$
$$\varepsilon_t = \sqrt{\sigma_t}z_t\sigma^2 = \omega + \sum_{i=1}^{p}\alpha_i\varepsilon_{t-i}^2 + \sum_{i=1}^{q}\beta_i\sigma_{t-i}^2 \quad (4)$$

where α_i, with $i = 1, ..., p$, and β_i, with $i = 1, ..., q$ 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:

$$f_X(x\mid\mu,\sigma,\nu) = \begin{cases} \nu, \text{if } x = 0 \\ (1-\nu)\, f_{GA}(x\mid\mu,\sigma), \text{if } x > 0 \end{cases} \quad (5)$$

For $0 \le x < \infty$, $0 < \nu < 1$, 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:

$$f_X(x\mid\mu,\sigma,\nu) = \begin{cases} \nu, \text{if } x = 0 \\ (1-\nu)\, f_{IG}(x\mid\mu,\sigma), \text{if } x > 0 \end{cases} \quad (6)$$

For $0 \le x < \infty$, $0 < \nu < 1$, 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 Equation (4), 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 Equation (4)).

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 $f(x_1, ..., x_d)$ of the random vector $\mathbf{X} = (X_1, ..., X_d)$ as a product of conditional densities:

$$f\left(x_1,\ldots,x_d\right) = f_d\left(x_d\right)\cdot f_{d-1d}\left(x_{d-1} \mid x_d\right)\cdot\ldots \cdot f_{1|2\ldots d}\left(x_1 \mid x_2,\ldots,x_d\right) \tag{7}$$

The factorisation in Equation (7) 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:

$$f_{d-1|d}\left(x_{d-1} \mid x_d\right) = c_{d-1,d}\left(F_{d-1}\left(x_{d-1}\right),F_d\left(x_d\right);\vartheta_{d-1,d}\right) \cdot f_{d-1}\left(x_{d-1}\right) \tag{8}$$

where $c_{d-1,d}$ is a bivariate copula, with parameter vector $\vartheta_{d-1,d}$. Through a straightforward generalisation of Equation (8), each term in Equation (7) 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:

$$f_{X_j|v\left(x_j|\mathbf{v}\right)} = c_{X_j,v_l;\mathbf{v}_{-l}}\left(F_{X_j|\mathbf{v}_{-l}}\left(X_j \mid \mathbf{v}_{-l}\right),F_{v_l|\mathbf{v}_{-l}}\left(v_l \mid \mathbf{v}_{-l}\right); \vartheta_{X_j,v_l;\mathbf{v}_{-l}}\right)\cdot f_{X_j|\mathbf{v}_{-l}}\left(X_j \mid \mathbf{v}_{-l}\right) \tag{9}$$

where $\mathbf{v}$ is the conditioning vector, v_l is a generic component of $\mathbf{v}$, $\mathbf{v}_{-l}$ is the vector $\mathbf{v}$ without the component v_l, $F_{X_j|\mathbf{v}_{-l}}\left(\cdot,\cdot\right)$ is the conditional distribution of X_j given $\mathbf{v}_{-1}$ and $c_{X_j,v_l;\mathbf{v}_{-l}}\left(\cdot,\cdot\right)$ is the conditional bivariate copula density, which can typically belong to any family (e.g. Gaussian, Student's *t*, Clayton, Gumbel, Frank, Joe, BB1, BB6, BB7, BB8, etc.; for more information on copula families, see Nelsen, 2007), with parameter $\vartheta_{X_j,v_l;\mathbf{v}_{-l}}\left(\cdot,\cdot\right)$. The *d*-dimensional joint multivariate distribution function can hence be expressed as a product of bivariate copulas and marginal distributions by recursively plugging Equation (9) in Equation (7).

For example, let us consider a 6-dimensional distribution. Then, Equation (7) translates to:

$$f\left(x_1,\ldots,x_6\right) = f_6\left(x_6\right)\cdot f_{5|6}\left(x_5 \mid x_6\right)\cdot f_{4|5,6}\left(x_4 \mid x_5,x_6\right)\cdot\ldots \cdot f_{1|2,\ldots,6}\left(x_1 \mid x_2,\ldots,x_6\right). \tag{10}$$

The second factor $f_{5|6}\left(x_5 \mid x_6\right)$ on the right-hand side of Equation (10) can be easily decomposed into the bivariate copula $c_{5|6}\left(F_5\left(x_5\right),F_6\left(x_6\right)\right)$ and marginal density $f_5\left(x_5\right)$:

$$f_{5|6}\left(x_5 \mid x_6\right) = c_{5,6}\left(F_5\left(x_5\right),F_6\left(x_6\right);\vartheta_{5,6}\right)\cdot f_5\left(x_5\right) \tag{11}$$

On the other hand, the third factor on the right-hand side of Equation (10) can be decomposed using the Equation (9) as:

$$f_{4|5,6}\left(x_4 \mid x_5,x_6\right) = c_{4,5;6}\left(F_{4|6}\left(x_4 \mid x_6\right),F_{5|6}\left(x_5 \mid x_6\right);\vartheta_{4,5;6}\right) \cdot f_{4|6}\left(x_4 \mid x_6\right) \tag{12}$$

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

$$f\left(x_1,\ldots,x_6\right) = \prod_{j=1}^{6} f_j\left(x_j\right)\cdot c_{1,2}\cdot c_{1,3}\cdot c_{3,4}\cdot c_{1,5}\cdot c_{5,6}\cdot c_{2,3;1} \cdot c_{1,4;3}\cdot c_{3,5;1}\cdot c_{1,6;5}\cdot c_{2,4;1;3}\cdot c_{4,5;1,3}\cdot c_{3,6;1,5} \cdot c_{2,5;1,3,4}\cdot c_{4,6;1,3,5}\cdot c_{2,6;1,3,4,5}. \tag{13}$$

Equation (13) is called pair copula construction. Note that in the previous equation the notation has been simplified, setting $c_{a,b} = c_{a,b}\left(F_a\left(x_a\right),F_b\left(x_b\right);\vartheta_{a,b}\right)$.

Two particular types of vines are the Gaussian vine and the Independence vine. The first one is constructed using solely Gaussian bivariate pair-copulas as building blocks, such that each conditional bivariate copula density $c_{X_j,v_l;\mathbf{v}_{-l}}\left(\cdot,\cdot\right)$ described in Equation (9) is a Gaussian copula. The Gaussian vine was adopted in flood risk analysis by Couasnon et al. (2018). The second type is the independence vine, which is constructed using only independence pair-copulas, that are simply given by the product of the marginal distributions of the random variables. In this latter case each conditional bivariate copula density $c_{X_j,v_l;\mathbf{v}_{-l}}\left(\cdot,\cdot\right)$ described in Equation (9) is an Independence copula, implying absence of dependence between the variables.

Pair copula constructions can be represented through a graphical model called regular vine (R-vine). An R-vine $V(d)$

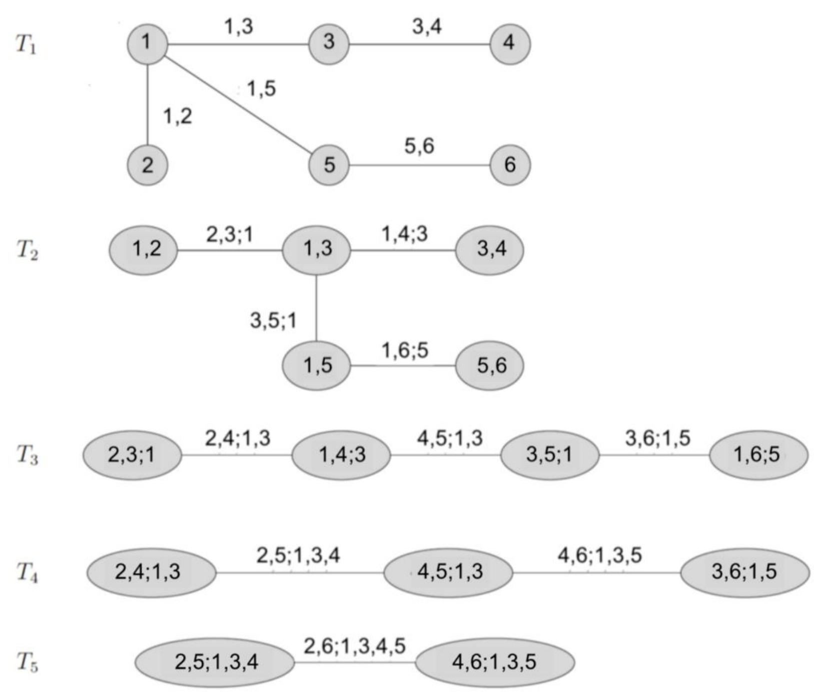

Figure 2. Six-dimensional R-vine graphical representation. Source: Czado (2019).

on d variables is a nested set of trees (connected acyclic graphs) T_1, ..., 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 $T_{k+1}, k = 1, \ldots, d-1$. Two edges can share a node in tree T_k without the associated nodes in tree T_{k+1} being connected. In an R vine, two edges in T_k which become two nodes in tree T_{k+1}, can only share an edge if in tree T_k the edges shared a common node, but they are not necessarily connected by an edge.

Figure 2 shows the 6-dimensional R-vine represented in Equation (13). 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). The copula parameters ϑ for each pair-copula in the vine are 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.

3.3. Out-of-Sample Predictions

In order to evaluate the suitability of the proposed vine copula model in relation to other methods, we produced one-day-ahead out-of-sample predictions and we compared them to the original data. 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 Remillard (2015). Before fitting the vine, we extracted the residuals from the marginals, as explained in Section 3.1, and obtained the *u-data*. Next, after fitting the vine, 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. More precisely, we applied the inverse transformation to the M realizations from the vine copula to obtain the residuals which we then plugged into the marginal models to get the predicted values of the environmental variables (wave height and water level). Then, we calculated the average prediction for all simulations $\mathbf{X}_{T+1}^{Avg}$ and use it as the forecast $\mathbf{X}_{T+1}$. The prediction interval of level $(1-\alpha) \in (0, 1)$ for 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 $\mathbf{X}^{l}_{T+1}$ and $\mathbf{X}^{u}_{T+1}$ the lower and upper values of the prediction intervals.

In order to compare and contrast the accuracy of predictions for different models, we made use of four indicators: the mean squared error (MSE) to evaluate point forecasts; the mean interval score (MIS), proposed by Gneiting and Raftery (2007), to assess the accuracy of the prediction intervals; the Normalized Nash-Sutcliffe model efficiency (NNSE) coefficient, proposed by Nash and Sutcliffe (1970) to appraise hydrological models; and the Distance Correlation, proposed by Szekely et al. (2007), to determine the association between observed and predicted data. The MSE for each variable $j = 1, ..., d$ was calculated as follows:

$$\mathrm{MSE}_j = \frac{1}{S}\sum_{t=T+1}^{T+S}\left(x_{t,j} - x_{t,j}\right)^2 \tag{14}$$

where $x_{t,j}$ is the observed value for each variable at each time point t, $x_{t,j}$ is the corresponding predicted value, $T+1$ denotes the first predicted date, while $T+S$ indicates the last predicted date. The 95% MIS for each variable, at level $\alpha = 0.05$, was computed as:

$$\mathrm{MIS}_j = \frac{1}{S}\sum_{t=T+1}^{T+S}\left[\left(x^u_{t,j} - x^l_{t,j}\right) + \frac{2}{\alpha}\left(x^l_{t,j} - x_{t,j}\right)1\left(x_{t,j} < x^l_{t,j}\right) + \frac{2}{\alpha}\left(x_{t,j} - x^u_{t,j}\right)1\left(x_{t,j} > x^u_{t,j}\right)\right] \tag{15}$$

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

The NNSE coefficient was calculated as:

$$\mathrm{NNSE}_j = 1/(2 - \mathrm{NSE}_j) \tag{16}$$

with

$$\mathrm{NSE}_j = 1 - \frac{\sum_{t=T+1}^{T+S}\left(x_{t,j} - x_{t,j}\right)^2}{\sum_{t=T+1}^{T+S}\left(x_{t,j} - \bar{x}_j\right)^2} \tag{17}$$

where $\bar{x}_j$ is the mean of the observed values for each variable. The NSE is a normalized statistic that determines the relative magnitude of the residual variance ("noise") compared to the measured data variance ("information"). The Distance Correlation takes the form:

$$\mathrm{DC}_j = \mathrm{dCor}\left(X_j, X_j\right) = \frac{\mathrm{dCov}\left(X_j, X_j\right)}{\sqrt{\mathrm{dVar}\left(X_j\right)\mathrm{dVar}\left(X_j\right)}} \tag{18}$$

where X_j is the j^{th} observed variable, X_j is the corresponding j^{th} predicted variable, $\mathrm{dCov}(X_j, X_j)$ is the distance covariance and $\mathrm{dVar}(X_j)$ and $\mathrm{dVar}(X_j)$ are the distance standard deviations, obtained replacing the signed distances between the variables with centred Euclidean distances. The DC is a distance-based correlation that can detect both linear and non-linear relationships between variables.

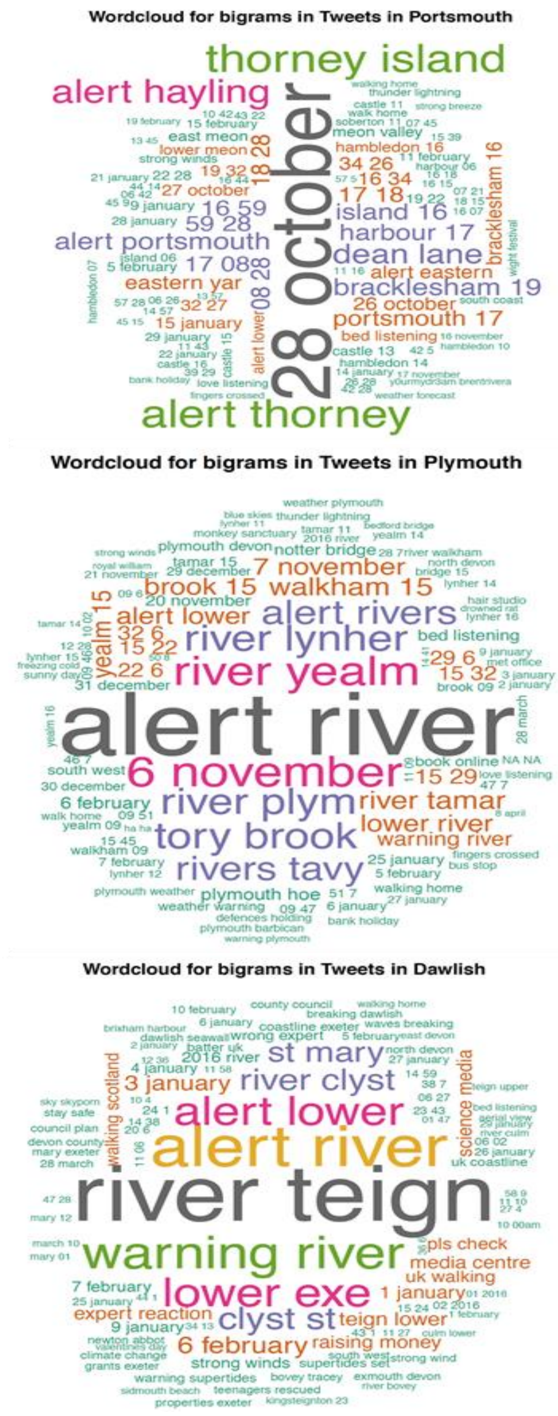

Figure 3. Wordclouds of paired words in tweets from Portsmouth (top panel), Plymouth (middle panel) and Dawlish (bottom panel).

4. Result Analysis and Discussions

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 3 displays the wordclouds of paired words obtained by pairing 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. 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).

4.2. Marginals Estimation

Table S1 (in the supplementary materials) 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.

As an example, Figure 4 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, the middle panel illustrates the observations (black line) and in-sample predictions obtained from the fitted ZAGA model (red line), while the bottom panel shows the histogram of the resulting *u-data*. 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, the in-sample predictions overlapping the observed data and the shape of the *u-data* histogram displaying a uniform pattern.

4.3. Vine Copula Estimation

Once the marginals were estimated, 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 5 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. 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 5, 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 contiguously related. The symmetric Gaussian copula, which is often employed in traditional multivariate modelling, was not identified as the best fitting copula for any 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.

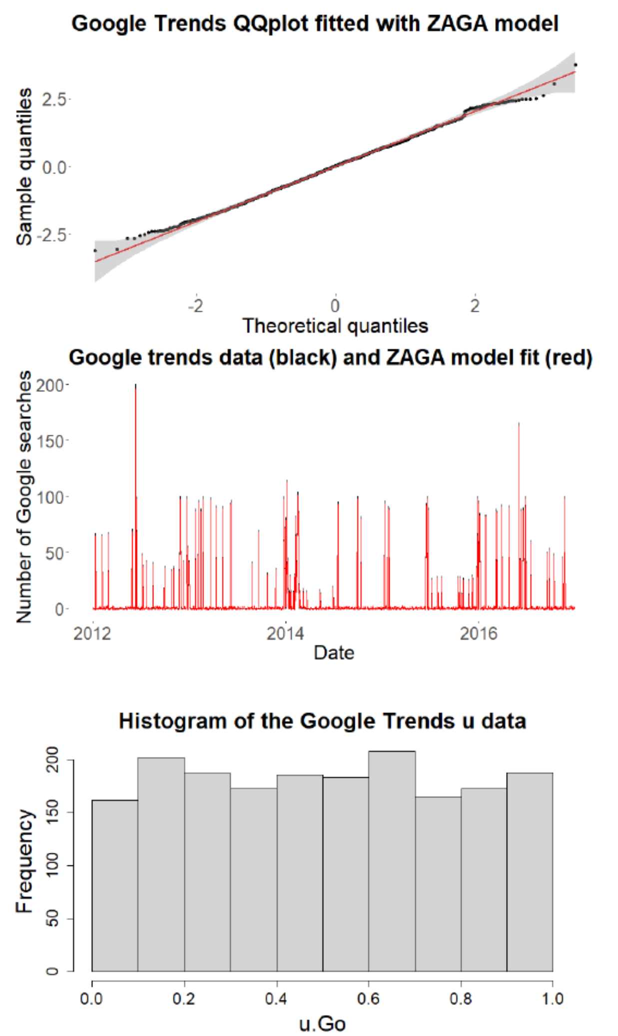

Figure 4. Plots illustrating the fit of the residuals for the Google marginal for Portsmouth. Top panel: QQ-plot comparing the Gaussian theoretical quantiles with sample quantiles. Middle panel: observed time series (black line) and in-sample predictions obtained from the fitted ZAGA model (red line). Bottom panel: Histogram of the resulting u-data.

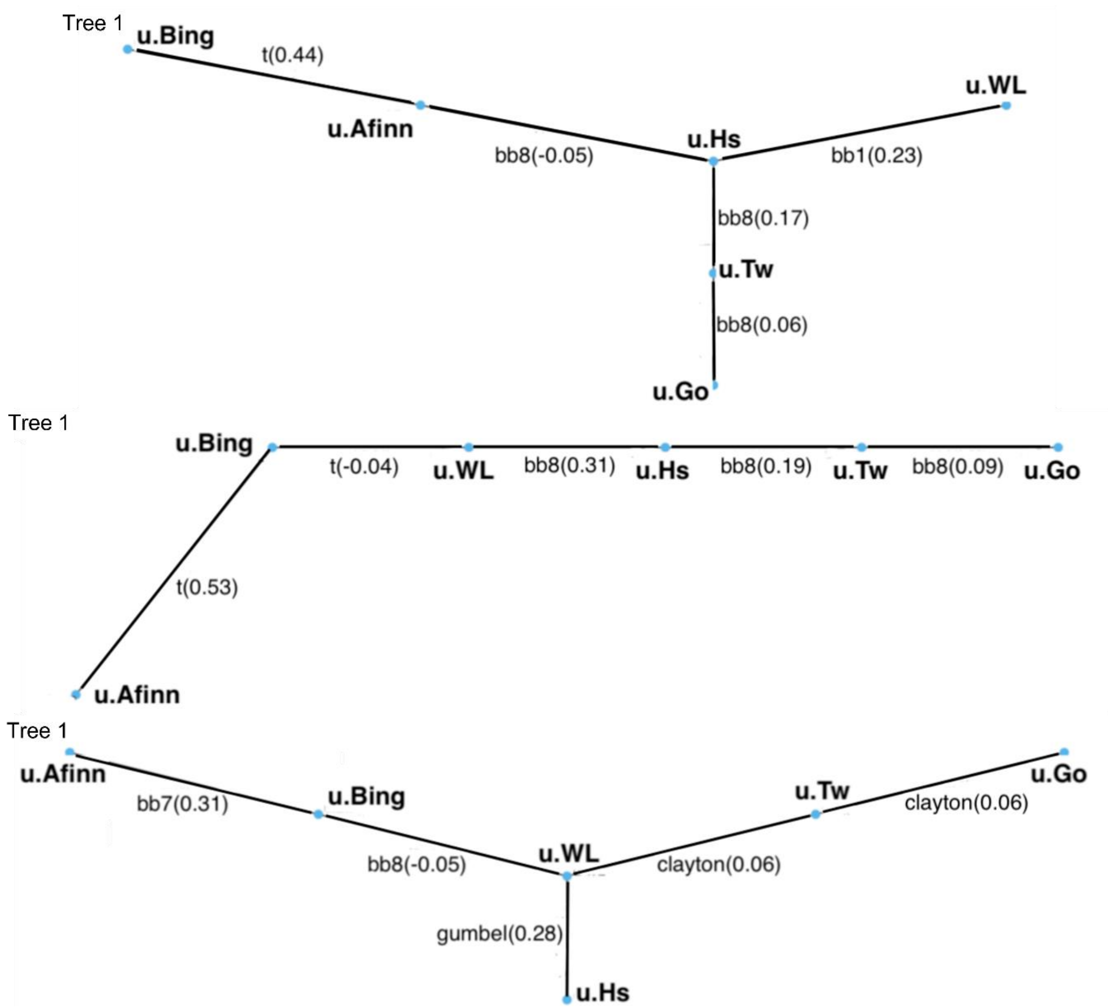

Figure 5. First trees of the vine copulas estimated for Portsmouth (top panel), Plymouth (middle panel) and Dawlish (bottom panel).

4.4. Out-of-Sample Prediction Results

In this Section we constructed out-of-sample predictions using the proposed vine methodology, which integrates environmental and social media variables. We then compared 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 as illustrated in Section 3.3, considering the vine copula estimated as explained 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.

Tables 2 and 3 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 results assuming independence among variables, the third columns show the results assuming all Gaussian pair-copulas, and the fourth columns show the vine copula results. The MSEs and MISs of the best performing approaches for each variable are highlighted in boldface. From Tables 2 and 3, we notice that the vine copula approach outperforms the other two approaches in the majority of the cases. 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 a few cases. 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 S2, making it difficult to define associations between online and environmental data and to leverage data integration for predicting purposes.

Table 2. MSEs Calculated for Portsmouth (Top Panel), Plymouth (Middle Panel) and Dawlish (Bottom Panel) for Each Variable

MSE Portsmouth			
Variable	Independent	Gaussian	Vine Copula
Hs	0.2693	**0.2603**	0.2639
WL	**0.0301**	0.0327	0.0325
Google	404.4304	**403.9977**	404.4147
Total Tweets	6.7351	**6.6994**	6.7829
Bing	**2.6572×10^{-11}**	2.6634×10^{-11}	2.6624×10^{-11}
Afinn	1.3823×10^{-10}	**1.3745×10^{-10}**	1.3767×10^{-10}
MSE Plymouth			
Variable	Independent	Gaussian	Vine Copula
Hs	0.3646	0.3647	**0.358**
WL	0.0274	0.0278	**0.0261**
Google	2874.761	2875.053	**2873.466**
Total Tweets	14.2388	14.1698	**14.1569**
Bing	2.6834×10^{-11}	**2.6282×10^{-11}**	2.6829×10^{-11}
Afinn	1.2103×10^{-10}	1.2035×10^{-10}	**1.2028×10^{-10}**
MSE Dawlish			
Variable	Independent	Gaussian	Vine Copula
Hs	0.2857	**0.2864**	0.2915
WL	**0.0267**	0.0295	0.0285
Google	4612.772	4613.572	**4612.738**
Total Tweets	**609.9969**	610.042	610.3111
Bing	5.7304×10^{-9}	**5.6124×10^{-9}**	5.6264×10^{-9}
Afinn	6.1873×10^{-9}	6.1670×10^{-9}	**6.1208×10^{-9}**

Note: The numbers show the results assuming independence among variables (second column), all gaussian pair-copulas (third column) and vine copula (fourth column); the mses of the best performing approaches for each variable are in boldface.

The variables Hs and WL are generally better predicted by the vine method, as opposed to the independence approach, which assumes no dependence between any of the variables involved in the model. Hence, the independence approach indicates the absence of any association between the environmental and the social media variables, implying the lack of contribution of online-generated information in predicting the flood variables. On the contrary, the vine approach assumes the presence of a dependence structure between the variables and, in particular, between the environmental and social media insights. Therefore, the better performance of the vine compared to the independence model demonstrates usefulness of social media information in forecasting environmental variables.

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 Total tweets achieves generally better results with the vine copula method for Plymouth data and with the Gaussian method for Portsmouth data, while the independence approach is typically selected for Dawlish data, due to the lack of information for this location, as explained above.

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.

Table 3. MISs Calculated for Portsmouth (Top Panel), Plymouth (Middle Panel) and Dawlish (Bottom Panel) for Each Variable

MIS Portsmouth			
Variable	Independent	Gaussian	Vine Copula
Hs	0.4193	0.4158	**0.1199**
WL	**0.0431**	0.0436	0.0465
Google	6.1179	**6.1151**	6.1169
Total Tweets	0.6366	**0.6316**	0.6356
Bing	1.2021×10^{-6}	**1.1982×10^{-6}**	1.2039×10^{-6}
Afinn	3.175×10^{-6}	3.1668×10^{-6}	**3.1644×10^{-6}**
MIS Plymouth			
Variable	Independent	Gaussian	Vine Copula
Hs	0.1554	0.1533	**0.1518**
WL	0.0384	0.0365	**0.0361**
Google	10.8789	10.879	**10.8759**
Total Tweets	0.7849	0.7845	**0.7833**
Bing	1.2178×10^{-6}	**1.2043×10^{-6}**	1.2175×10^{-6}
Afinn	3.0799×10^{-6}	**3.0693×10^{-6}**	3.0704×10^{-6}
MIS Dawlish			
Variable	Independent	Gaussian	Vine Copula
Hs	0.4177	**0.4116**	0.4116
WL	**0.0383**	0.0388	0.0431
Google	13.5782	13.5794	**13.5782**
Total Tweets	**7.0887**	7.0889	7.0913
Bing	1.7472×10^{-5}	**1.7184×10^{-5}**	1.7284×10^{-5}
Afinn	1.8396×10^{-5}	1.8385×10^{-5}	**1.8301×10^{-5}**

Note: The numbers show the results assuming independence among variables (second column), all gaussian pair-copulas (third column) and vine copula (fourth column); the miss of the best performing approaches for each variable are in boldface.

Figure S3 (in the supplementary materials) depicts grouped bar charts showing the differences between optimal fit for each model and the NNSEs for wave height (left panel) and water level (right panel) for each location. The red bars show the results assuming independence among variables, the green bars assuming all Gaussian pair-copulas and the blue bars assuming a vine copula model. Shorter bars indicate better fitting models. In the Plymouth location, the vine copula achieves better results than the other two models for both Hs and WL. The Gaussian model performs best for Hs in the Portsmouth location. The independent model is selected for the remaining cases, particularly in Dawlish, where again the lack of data points might be the cause.

Figure 6 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 ob-

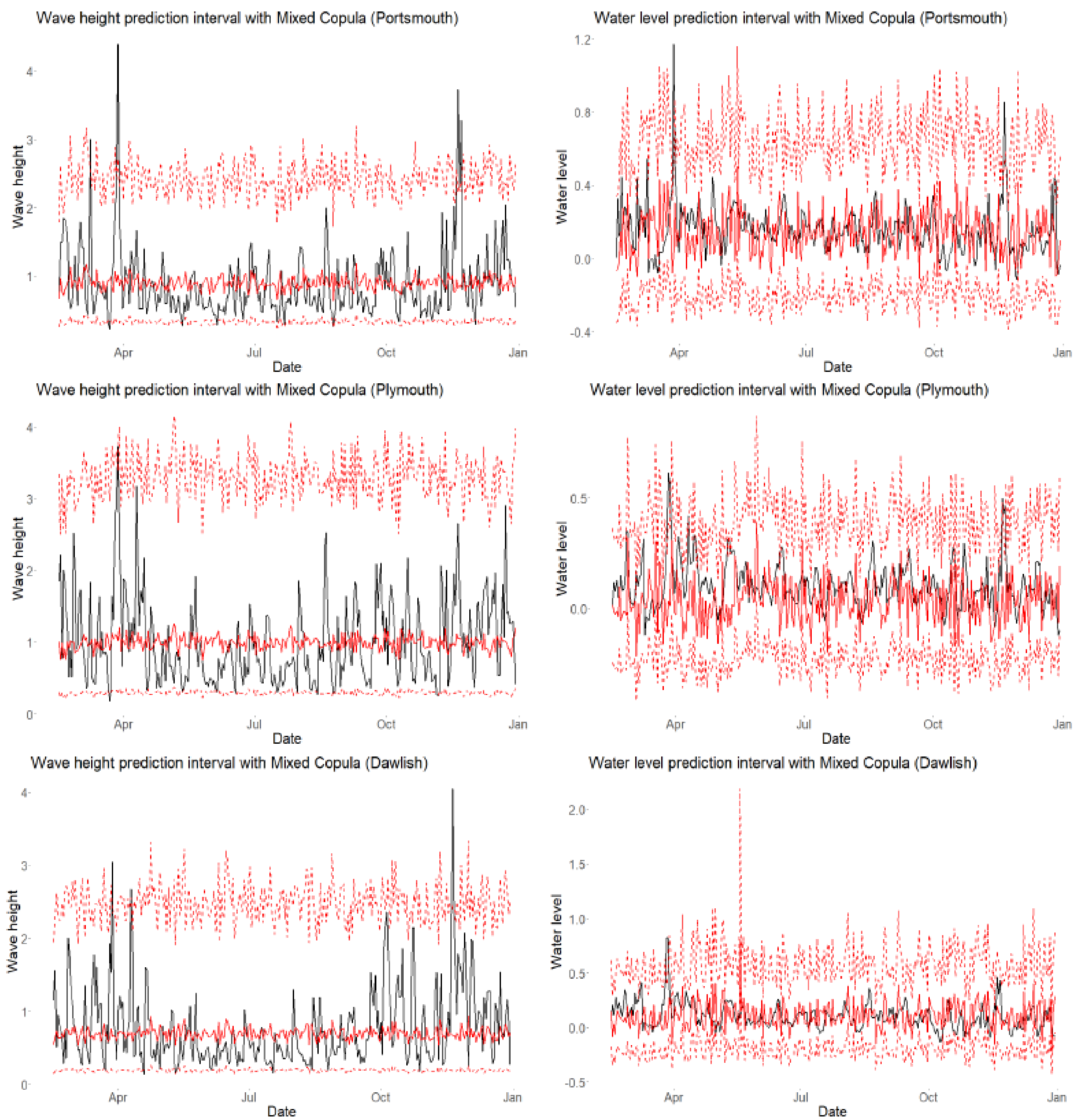

Figure 6. Line plots showing forecasts and prediction intervals for Hs (left panels) and WL (right panel) obtained with the vine copula methodology for the period between the 16th February 2016 and the 31st December 2016, for Portsmouth (top panel), Plymouth (middle panel) and Dawlish (bottom panel). Observed values are in black, predicted values are the inner red lines and 95% prediction intervals are the outer dotted red lines.

served values, the inner red lines denote the predicted values and the outer dotted red lines denote the 95% prediction intervals. We notice that the forecasted water levels are in line with the observations, and the average dynamics of wave height is adequately represented by the proposed model. Intervals predicted by 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 for forecasting flood-related data.

In addition, we carried out a correlation analysis between predicted and observed data. Figure S4 (in the supplementary materials) illustrates grouped bar charts showing the differences between optimal fit for each model and the DCs for wave height (left panel) and water level (right panel) for each location. The bar colour codes are the same used in Figure S3. According to the DC, the Gaussian vine model is generally the preferred approach, while the vine copula model performs best for wave height in the Plymouth location. The independent model, which implies no input from the social media data for calculating predictions, is never selected.

5. Concluding Remarks

In this paper, we propose a new methodology aimed at obtaining more accurate forecasts, compared to traditional approaches, for 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 obtained data at three different locations on the South coast of the UK, which were affected by severe storm events on several occasions in the past few years. Together with wave height and water level information, we also gathered Google Trends searches and Twitter microblogging 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. 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 than the other two approaches. 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.

The proposed methodology will support decision-makers enabling them to use knowledge gained from the model results to deepen their understanding of risks associated to floods and optimise resources in a more effective and efficient way. At strategic level, the methodology could be used to validate resource deployments in response to threats from floods; while at operational level, the methodology could assist to improve the effectiveness of civil contingency responses to flood events.

Further investigations involving other locations and including additional social media information will be the object of future work. Also, we will explore the use of the results of the study to validate inundation modes. Another extension will involve Bayesian inference, which would allow us to incurporate other information, such as experts' opinion, in the model. In addition, the use of more sophisticated machine learning approaches could be envisaged for deriving the sentiment variables to improve the proposed methodology.

Acknowledgments. The authors are grateful to the anonymous Reviewers for their useful comments which significantly improved the quality of the paper. 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., Czado, C., Frigessi, A. and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insur. Math. Econ*., 44(2), 182-198. https://doi.org/10.1016/j.insmatheco.2007.02.001

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

Arthur, R., Boulton, C.A., Shotton, H. and Williams, H.T. (2018). Social sensing of floods in the uk. *PloS one*. 13(1), e0189327. https://doi.org/10.1371/journal.pone.0189327

Balogun, A., Quan, S., Pradhan, B., Dano, U. and Yekeen, S. (2020). An improved flood susceptibility model for assessing the correlation of flood hazard and property prices using geospatial technology and fuzzy-ANP. *J. Environ. Inform*., https://doi.org/10.3808/jei.202000442

Bevacqua, E., Maraun, D., Hobæk Haff, I., Widmann, M. and Vrac, M. (2017) Multivariate statistical modelling of compound events via pair-copula constructions: analysis of floods in Ravenna (Italy), *Hydrol. Earth Syst. Sci*., 21, 2701-2723. https://doi.org/10.5194/hess-21-2701-2017

Brouwer, T., Eilander, D., Loenen, A.v., Booij, M.J., Wijnberg, K.M., Verkade, J.S. and Wagemaker, J. (2017). Probabilistic flood extent estimates from social media flood observations. *Nat. Hazards Earth Syst. Sci*., 17(5), 735-747. https://doi.org/10.5194/nhess-17-735-2017

Couasnon, A., Sebastian, A. and Morales-Nápoles, O. (2018). A copula-based Bayesian network for modeling compound flood hazard from riverine and coastal interactions at the catchment dcale: An application to the Houston Ship Channel, Texas, *Water*, 10(9), 1190. https://doi.org/10.20944/preprints201808.0072.v4

Czado, C. (2019). *Analysing dependent data with vine copulas*. Lecture Notes in Statistics, Springer. https://doi.org/10.1007/978-3-030-13785-4

De Albuquerque, J.P., Herfort, B., Brenning, A. and A. Zipf (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. Int. *J. Geogr. Inf. Sci*., 29(4), 667-689. https://doi.org/10.1080/13658816.2014.996567

Dissmann, J., Brechmann, E.C., Czado, C. and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Comput. Stat. Data. Anal.*, 59, 52-69. https://doi.org/10.1016/j.csda.2012.08.010

Feng, Y., Shi, P., Qu, S., Mou, S., Chen, C. and Dong, F. (2020). Non-stationary flood coincidence risk analysis using time-varying copula functions. *Sci. Rep*., 10(1), 1-12. https://doi.org/10.1038/ s41598-020-60264-3

Field, C.B., Barros, V., Stocker, T.F. and Dahe, Q. (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. https://doi. org/10.1017/CBO9781139177245

Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc*., 102(477), 359-378. https://doi.org/10.1198/016214506000001437

Grego, J.M., Yates, P.A. and Mai, K. (2015). Standard error estimation for mixed flood distributions with historic maxima. *Environmetrics* 26(3), 229-242. https://doi.org/10.1002/env.2333

Heffernan, J.E. and Tawn, J.A. (2004). A conditional approach for multivariate extreme values (with discussion). *J. R. Stat. Soc. Series. B Stat. Methodol.*, 66(3), 497-546. https://doi.org/10.1111/j.1467-9868.2004.02050.x

Herfort, B., de Albuquerque, J.P., Schelhorn, S.-J. and Zipf, A. (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. https://doi.org/10.1007/978-3-319-03611-3_4

Hu, M. and Liu, B. (2004). *Mining and summarizing customer reviews.* In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168-177. https://doi.org/10.1145/1014052.1014073

Huang, K., Dai, L., Yao, M., Fan, Y. and Kong, X. (2017). Modelling dependence between traffic noise and traffic flow through an entropy-copula method. *J. Environ. Inform*., 29(2). https://doi.org/10.3808/jei.201500302

Hyndman, R.J. and Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts. https://otexts.com/fpp3/

Jane, R., Dalla Valle, L., Simmonds, D. and Raby, A. (2016). A copula-based approach for the estimation of wave height records through spatial correlation. *Coast. Eng.*, 117, 1-18. https://doi.org/10.1016/j.coastaleng.2016.06.008

Jane, R.A., Simmonds, D.J. Gouldby, B.P., Simm, J.D., Dalla Valle, L. and Raby, A.C. (2018). Exploring the potential for multivariate fragility representations to alter flood risk estimates. *Risk Anal.*, 38(9), 1847-1870. https://doi.org/10.1111/risa.13007

Joe, H. (1997). *Multivariate models and multivariate dependence concepts.* CRC Press. https://doi.org/10.1201/9780367803896

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

Keef, C., Tawn, J.A. and Lamb, R. (2013). Estimating the probability of widespread flood events. *Environmetrics* 24(1), 13-21. https://doi.org/10.1002/env.2190

Latif, S. and Mustafa, F. (2020). Parametric vine copula construction for flood analysis for Kelantan river basin in Malaysia. *Civ. Eng. J.*, 6(8), 1470-1491. https://doi.org/10.28991/cej-2020-03091561

Li, Z., Wang, C., Emrich, C.T. and Guo, D. (2018). A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 South Carolina floods. *Cartogr. Geogr. Inf. Sci.*, 45(2), 97-110. https://doi.org/10.1080/15230406.2016.1271356

Mason, D.C., Davenport, I.J., Neal, J.C., Schumann, G.J.P., and Bates, P.D. (2012). Near real-time flood detection in urban and rural areas using high-resolution synthetic aperture radar images. *IEEE Trans. Geosci. Remote. Sens.*, 50(8), 3041-3052. https://doi.org/10.1109/TGRS.2011.2178030

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

Moishin, M., Deo, R.C., Prasad, R. Raj, N. and Abdulla, S. (2020). Development of flood monitoring index for daily flood risk evaluation: case studies in Fiji. *Stoch. Environ. Res. Risk Assess*., 1-16. https://doi.org/10.1007/s00477-020-01899-6

Muller, C., Chapman, L. Johnston, S. Kidd, C. Illingworth, S. Foody, G. Overeem, A. and R. Leigh (2015). Crowdsourcing for climate and atmospheric sciences: Current status and future potential. *Int. J. Climatol*., 35(11), 3185-3203. https://doi.org/10.1002/joc.4210

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

Nash, J.E. and Sutcliffe, J.V. (1970). River flow forecasting through conceptual models part I - A discussion of principles, *J. Hydrol*., 10(3), 282-290. https://doi.org/10.1016/0022-1694(70)90255-6

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 Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 54(3), 507-554. https://doi.org/10.1111/j.1467-9876.2005.00510.x

Rigby, R.A., Stasinopoulos, M.D., Heller, G.Z. and De Bastiani, F. (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R.* CRC press. https://doi.org/10.1201/9780429298547

Rosser, J.F., Leibovici, D. and Jackson, M. (2017). Rapid flood inundation mapping using social media, remote sensing and topographic data. *Nat. Hazards*. 87(1), 103-120. https://doi.org/10.1007/s11069-017-2755-0

Santos, V.M., Wahl, T., Jane, R., Misra, S.K. and White, K.D. (2021) Assessing compound flooding potential with multivariate statistical models in a complex estuarine system under data constraints. *J. Flood Risk Manag.*, e12749 https://doi.org/10.1111/ jfr3.12749

Saravanou, A., Valkanas, G., Gunopulos, D. and Andrienko, G. (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. https://doi.org/10.1145/2740908.2741730

Schumann, G., Bates, P.D., Horritt, M.S., Matgen, P. and Pappenberger, F. (2009). Progress in integration of remote sensing-derived flood extent and stage data and hydraulic models. *Rev. Geophys.*, 47(4). https://doi.org/10.1029/2008RG000274

Sibley, A., Cox, D. and Titley, H. (2015). Coastal flooding in England and Wales from Atlantic and North Sea storms during the 2013/2014 winter. *Weather*, 70(2), 62-70. https://doi.org/10.1002/wea.2471

Silge, J. and D. Robinson (2016). tidytext: Text mining and analysis using tidy data principles in R. *J. Stat. Softw.*, 1(3). https://doi.org/10.21105/joss.00037

Simard, C. and Remillard, B. (2015). Forecasting time series with multivariate copulas. *Dependence Model.*, 3(1). https://doi.org/10.1515/demo-2015-0005

Sklar, M. (1959). *Fonctions de repartition a n dimensions et leurs marges.* Publications de l'Institut de Statistique de l'Universite de Paris 8, 229-231.

Smith, L., Liang, Q. James, P. and Lin, W. (2017). Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *J. Flood Risk Manag*., 10(3), 370-380. https://doi.org/10.1111/jfr3.12154

Spielhofer, T., Greenlaw, R. Markham, D. and Hahne, A. (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), 1-6. IEEE. https://doi.org/10.1109/ICT-DM.2016.7857213

Spruce, M.D., Arthur, R., Robbins, J. and Williams, H.T. (2021). Social sensing of high-impact rainfall events worldwide: A benchmark comparison against manually curated impact observations. *Nat. Hazards Earth Syst. Sci.*, 1-31. https://doi.org/10.5194/nhess-2020-413

Szekely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007). Measuring and Testing Dependence by Correlation of Distances, *Ann. Stat*., 35(6), 2769-2794. https://doi.org/10.1214/009053607000000505

Talukdar, S., Ghose, B. Salam, R., Mahato, S., Pham, Q.B., Linh, N.T.T., Costache, R. and Avand, M. (2020). Flood susceptibility modeling in Teesta river basin, Bangladesh using novel ensembles of bagging algorithms. *Stoch. Environ. Res. Risk Assess*., 34(12), 2277-2300. https://doi.org/10.1007/s00477-020-01862-5

Tosunoglu, F., Gurbuz, F. and Ispirli, M.N. (2020). Multivariate mod - eling of flood˙ characteristics using vine copulas. *Environ. Earth Sci.*, 79(19), 1-21. https://doi.org/10.1007/s12665-020-09199-6

UN (2015). *The human cost of weather related disasters 1995-2015*, United Nations, Geneva, Switzerland.

Wang, X. and Du, C. (2003). An internet based flood warning system. *J. Environ. Inform*., 2(1), 48-56. https://doi.org/10.3808/jei.200300017

Ward, P.J., Couasnon, A., Eilander, D., Haigh, I.D., Hendry, A., Muis, S., Veldkamp, T.I.E., Winsemius, H.C. and Wahl, T. (2018). Dependence between high sea-level and high river discharge increases flood hazard in global deltas and estuaries, *Environ. Res. Lett.*, 13(8), 084012. https://doi.org/10.1088/1748-9326/aad400