The Advanced Meteorology Explorer: a novel stochastic, gridded daily rainfall generator

Laura C. Dawkins[1,*], **Joe M. Osborne**[1], **Theodoros Economou**[2], **Geoff J. C. Darch**[3], and **Oliver R. Stoner**[4]

[1]Met Office, Fitzroy Road, Exeter, Devon, EX1 3PB, UK

[2]Climate and Atmosphere Research Center, The Cyprus Institute, Cyprus

[3]Anglian Water Services, Cambridgeshire, UK

[4]School of Mathematics and Statistics, University of Glasgow, Glasgow, UK

[*]laura.dawkins@metoffice.gov.uk

Abstract

Synthetic rainfall simulations from stochastic models are commonly used for water resource management, as they are able to provide a wider range of meteorological conditions than those seen in the observed record. Here, we present a novel stochastic rainfall modelling framework, the Advanced Meteorology Explorer (AME), which combines and extends existing methods to enhance model flexibility, and meet a number of key water industry needs. This framework allows for the simulation of physically consistent synthetic daily rainfall data, coherently in space and time, on a high-resolution grid over a region of interest. The AME uses an advanced hidden Markov model structure within a Bayesian hierarchical framework to represent daily rainfall at a set of locations in a region, conditional on important climate drivers. The climate drivers included in the rainfall model at each location are able to vary using penalised regression, ensuring a transferable model that can be applied to different locations without adaptation. The dependence between locations is modelled following a flexible copula approach, able to capture varying dependence structures within the data, allowing for spatially coherent simulations at the modelled locations. Simulations are then interpolated to a high-resolution grid using a terrain adjusted inverse-distance weighted interpolation method.

The AME framework is applied to 105 years (1914-2018) of daily rainfall data at 39 sites in the Greater Anglian region

of the UK, and used to generate 1000 alternative realisations of the same period on a 5 km grid over the region. Validation of these simulations shows how the AME framework is able to accurately capture rainfall occurrence and intensity, as well as long-duration meteorological drought behaviour, important for quantifying water resource risk in this dry region. This framework has the potential to be applied to other regions, incorporate additional weather variables and the effect of climate change, and the resulting simulations can be used for environmental risk assessment in any industry impacted by rainfall.

Keywords: Stochastic rainfall model; Hidden Markov Model; Bayesian hierarchical framework; Copula; Vine-copula; Penalised Regression; Meteorological drought characterisation.

1 Introduction

Stochastic models of multivariate rainfall behaviour have been used within the field of water resource management for many years (Wilks and Wilby 1999; Serinaldi and Kilsby 2012; Breinl et al. 2013; Serinaldi and Kilsby 2014b; Verdin et al. 2019). These models simulate synthetic rainfall time series that can be used for environmental risk assessment. The intention is that these stochastic simulations capture a greater range of plausible meteorological scenarios than seen in the historical record, allowing for a more comprehensive quantification of possible water resource challenges. In particular, in the Greater Anglian region of the UK, stochastic rainfall simulations are commonly used to explore the range of potential long-duration drought behaviours, as these are known to have a large impact on water resource management in this dry region. This insight allows for more informed infrastructure planning and decision making, ensuring the water industry can provide a resilient water service to those living in their region.

Specifically, the UK water industry has been using rainfall generators for water resources planning for a decade. In 2011 work was undertaken by Southern Water and its consultant Atkins, working with Newcastle University and the University of East Anglia. The rainfall generator produced from this (Serinaldi and Kilsby, 2012) was spatially coherent across Southern Water's three geographically discrete supply areas. The model was based on non-parametric bootstrap re-sampling and a parametric Generalised Additive Model for Location, Scale and Shape (GAMLSS). It accounted for the relationship between rainfall and circulation indices: the North Atlantic Oscillation (NAO) and sea surface temperature (SST). The model successfully simulated low (and high) rainfall more extreme than observed events, thus facilitating use in assessing the impact of severe drought.

This rainfall generator was adopted in practice within the UK water industry, with further work undertaken by industry to produce spatial, daily data for use in hydrological modelling. However, there were some limitations with the outcomes of this latter process (related to the industry application rather than the underlying modelling framework). Firstly, the generated hydrological droughts at the severe end of the scale (quantified using hydrological models applied to the simulated rainfall) tended to have a small but systematic wet skew when compared to available observations. In practice this was found to lead to a significant under preparedness for severe droughts. This necessitated a post-processing step which brought the median of the stochastic rainfall distribution close to the corresponding observed quantiles. Secondly, the need for daily catchment data meant that further data processing was required, adding uncertainty. The rainfall generator and post-processing steps were subsequently improved and added to, including development of a fully parametric version of the model, with application to water resources planning at company, regional and national levels (Water UK, 2016).

The rainfall generator was reviewed and updated by the Met Office in 2016 for use in the Water Resources East project led by Anglian Water. The rainfall generator review explored training the model on seasonal rather than monthly data, using monthly interdependencies on NAO and SST, and the addition of the East Atlantic (EA) pattern. The updated rainfall generator was applied to 39 sites across the east of England and used in hydrological modelling. Although reduced, the wet skew in the simulations remained. Even after the rainfall correction, flows were found to be skewed, perhaps due to the simplicity of the correction or due to issues in the post-processing steps such as the conversion to catchment averages.

To improve upon this current industry approach, UK water companies require a rainfall model that is based on sound statistical theory, and enables the simulation of synthetic rainfall time series with the following properties:

1. Simulated on a daily temporal resolution, to provide the granularity required for water resource management decision making, with the appropriate dependency in time;
2. Accurately capturing known behaviour of UK rainfall variability, such as that related to the effect of climate drivers;
3. On a high-resolution grid covering their water resource region, with appropriate dependency in space;
4. Providing realistic long-duration meteorological drought characteristics, to ensure water resource risk is appropriately quantified.

A wealth of stochastic rainfall and weather generators have been developed in recent years. These exist on monthly, daily, and sub-daily time-scales and have been developed with different levels of complexity and refinement to address different needs. Indeed, many examples provide realistic stochastic daily rainfall simulations at multiple locations, developed using a variety of approaches. These include non-parametric approaches, such as nearest neighbour bootstrapping (Buishand and Brandsma, 2001); semi-parametric approaches, where daily precipitation amounts are generated by first re-sampling observed values and then sampling and reshuffling synthetic precipitation amounts from parametric distribution functions (Breinl et al., 2013, 2015); and parametric models, which represent daily precipitation at individual locations using a chain-dependent process (Katz, 1977), and the spatial dependence between locations using latent Gaussian Processes (GPs) (Wilks, 2009; Kleiber et al., 2012; Serinaldi and Kilsby, 2014b; Oriani et al., 2018; Verdin et al., 2019). These parametric methods often model daily precipitation occurrence using a two-state, first-order Markov chain, and the (non-zero) precipitation amounts are drawn from a fitted probability distribution conditional on the simulation of a wet day. Serinaldi and Kilsby (2014b) show how using this first-order Markov chain model structure to simulate daily synthetic rainfall provides reasonable agreement with the observed lag-1, lag-2, lag-180 and lag-365 day temporal autocorrelations (relevant for addressing point 1 above). Moreover, daily rainfall extremes have been shown to be modelled well, especially when marginal distribution families are wisely selected, such as those described by Li et al. (2012) and Naveau et al. (2016).

Parametric Hidden Markov Models (HMMs) have also been developed for multi-site daily stochastic rainfall simulation (Holsclaw et al., 2016; Kroiz et al., 2020). HMMs characterise rainfall as being associated with one of a finite number of 'hidden' states (e.g. dry, wet and very wet), and model rainfall magnitude and occurrence in each state separately using a different statistical distribution. The HMM therefore consists of two parts, firstly a model for the transitions between states at consecutive time steps (e.g. the probability of transitioning into a wet state in the next time-step, given you are currently in a dry state). This is also often represented using a first-order Markov chain. Secondly, within each of the HMM states, a two stage model for the occurrence of rainfall and, if rainfall occurs, the quantity of rainfall at that given time-step. This hidden state structure provides a great deal of flexibility, allowing for temporal structures in both the rainfall occurrences and intensities to be captured together, through the persistence of each state, meaning higher temporal resolutions can be modelled with increased accuracy (Stoner and Economou, 2020a). Specifically, Stoner and Economou (2020a) show how a single site, three state, HMM is generally able to capture the temporal dependency structure of hourly rainfall data well (again, relevant for addressing point 1 above).

A further advantage of these parametric approaches is in their ability to introduce non-stationarity within the model parameters, through the influence of time-varying covariates, such as large-scale atmospheric indicators or climate drivers (as required by point 2 above). This has been found to be important for preserving the long-term variability of stochastic rainfall simulations, particularly the length of both the dry and wet extreme spells (Oriani et al., 2018), therefore also relevant for point 4 above. The underestimation of the true natural variability by weather generators is typically indicated as an "overdispersion" problem (e.g., Katz and Parlange 1998). While the introduction of time-varying covariates may address this problem, it should be noted that this may not necessarily be for the right reason, as overdispersion could be partially related to the stochastic generator structure rather than to non-stationarity. In existing models, non-stationarity is most commonly achieved using generalised linear models (GLM) or GAMLSS frameworks, which capture how time-varying covariates influence the occurrence and magnitude of rainfall on each day. For example, Furrer and Katz (2007) use a GLM approach to model and simulate rainfall in the Pampas region of central-eastern Argentina, including within their models for rainfall occurrence and intensity the influence of the El Niño-Southern Oscillation (ENSO) index and time. Similarly, Serinaldi and Kilsby (2014b) use a GAMLSS approach to simulate realistic rainfall fields at a daily time scale over the Danube basin, conditioning their models for rainfall occurrence and intensity on sea level pressure and time. Moreover, Lister et al. (2018) suggest that using stronger predictor-precipitation relationships could further improve the performance of rainfall generators.

A number of these parametric models have also been developed to simulate rainfall on a high-resolution spatial grid (point 3 above). Most commonly, this is achieved using latent spatial GPs. Examples in the literature use such models to represent the gridded spatial structure of the weather generator model parameters themselves (such as the probability of rainfall occurrence) and/or the probability integral transformed 'residual' rainfall (Wilks, 2008; Verdin et al., 2015; Kleiber et al., 2012; Serinaldi and Kilsby, 2014b), similar to the concept of copula modelling (Nelson, 2006). The weather generator parameters are then modelled as a function of space, e.g. using locally weighted regressions (Wilks, 2008). By using a latent spatial GP, which captures the spatial dependence between two locations based on a correlation function of distance, the models are able to simulate spatially coherent rainfall at any desired location (i.e. a grid) within the modelled domain. Similar to this, in multi-site (non-gridded) examples, Gaussian copulas or multivariate Normal distributions are often used to capture the dependence between locations (Serinaldi and Kilsby, 2012; Kroiz et al., 2020; Chandler, 2019). A possible limitation of such methods is their underlying assumption of Gaussianity in the dependence structure. The Gaussian dependence structure assumes asymptotic extremal *independence* (Coles et al., 1999) between modelled variables (i.e. that the most extreme values tend not to occur together), hence the alternative, asymptotic extremal *dependence*, is not captured.

Here, we develop a novel stochastic rainfall modelling framework, henceforth referred to as the Advanced Meteorology Explorer (AME). The AME draws upon many of the advantages of the modelling frameworks referenced above to meet the needs of the UK water industry, extending them in a number of ways to enhance model flexibility. In particular, the AME framework is developed to closely match the observed rainfall behaviour within the target region. This is done to ensure that the subsequent simulations from the AME do not lead to water industry under preparedness for severe droughts, as was the case in previous models. To achieve this, the AME framework requires a larger number of parameters and several methodological solutions that increase complexity.

Within this paper, the AME modelling framework is described in Section 2. This framework is then applied to 105 years (1914-2018) of daily rainfall data at 39 sites in the Greater Anglian region, and used to generate 1000 alternative realisations of the same 105-year period on a 5 km grid over the region. Section 3 introduces the case study data, and the case study specific model set up is presented in Section 4. The performance of the model is then evaluated in Section 5, with particular focus on the representation of long-duration meteorological drought characteristics (relevant for addressing the water industry need in point 4 above). Finally, Section 6 discusses the model's strengths, weaknesses and potential areas of further development, and concludes.

2 Methodology

This section presents the stochastic rainfall modelling framework of the AME. The way in which the framework is specifically applied to model and simulate gridded daily rainfall in the Greater Anglian region of the UK is then presented in Section 4.

2.1 Overview of the AME framework

As described in Section 1, the intention is for the AME framework to closely capture the observed rainfall behaviour within a given region and period, to facilitate water resource management decision making. The intention is therefore to heavily depend on external covariates and the computation of time variable parameters, to allow for specific periods within the observed record (particularly extreme droughts) to be reproduced.

The AME follows a parametric framework in which the marginal distribution of rainfall at a given location is represented using an extension of the advanced HMM recently developed by Stoner and Economou (2020a). This HMM includes both temporally non-homogeneous state persistence probabilities, rainfall occurrence probabilities and rainfall distribution parameters, as well as 'clone' dry states. These clone states have identical rainfall distributions, but differing persistence probabilities, providing increased state transition flexibility through explicitly representing differing dry state persistence characteristics. See Stoner and Economou (2020a) and references therein for a more detailed description of the clone dry state modelling approach. Stoner and Economou (2020a) show how this flexible advanced HMM structure is able to capture well the distribution of dry period lengths; seasonal and annual variation in occurrence and intensity of rainfall (including extreme values); and the distribution of rainfall intensity when aggregated to daily and monthly resolutions, important for addressing the water industry need in point 1 above.

The AME builds upon the advanced HMM of Stoner and Economou (2020a) to include the influence of large-scale atmospheric indicators (climate drivers) within the rainfall occurrence and intensity models in each hidden state, hence addressing point 2 in Section 1. This is similar to the GLM-HMM model of Holsclaw et al. (2016), but with the additional flexibility facilitated by the inclusion of clone dry states and temporally non-homogeneous state persistence probabilities. Further model flexibility is introduced within the AME through the implementation of penalised regression (van Erp et al., 2019). Rather than fitting the same HMM to each location, based on the same climate drivers, as is commonly done (Furrer and Katz, 2007; Verdin et al., 2018), or carrying out a time-consuming variable selection step before model fitting, the in-built penalised regression within the AME framework allows for variable selection at the same time as model fitting. This provides a transferable model structure that can be fit to different sites/locations with no modification or prior model exploration. This additional flexibility is useful when modelling UK rainfall, because the relationship between rainfall and climate drivers is known to vary across this region (Lister et al., 2018). In addition, by 'shrinking' away non-influential climate drivers, the overall uncertainty of the HMM is reduced, hence improving model performance (van Erp et al., 2019). Moreover, within our application of the AME framework we present a method for tailoring the climate drivers to best capture their influence on rainfall within the modelled region, improving the representation of climate variability, as suggested by Lister et al. (2018).

Further, the AME extends the advanced HMM of Stoner and Economou (2020a) to allow for multi-site modelling, and subsequently gridded stochastic simulations to meet the water industry need described in point 3 in Section 1. Similar to a number of examples referenced above (Verdin et al., 2015; Kleiber et al., 2012; Serinaldi and Kilsby, 2014b), the site specific rainfall models are used to transform the daily rainfall at each site to standardised margins (using the probability integral transform), and the spatial dependence in the resulting residual rainfall is then modelled. Rather than using a latent GP, which assumes a Gaussian dependence structure, a more flexible multidimensional copula model is employed. In particular, the AME framework uses a vine-copula approach (Czado, 2019), able to capture differing (extremal) dependence structures between different pairs of modelled sites. This is found to be important when applying the AME framework to real world data. Stochastic simulation from this copula dependence model, combined with the individual site HMMs, provides spatially and temporally coherent synthetic daily rainfall time series at the modelled sites, capturing the effect of climate driver variables. The AME framework then interpolates these simulations to a high-resolution (5 km) spatial grid using the spatial interpolation method developed by the Met Office National Climate Information Centre (NCIC), previously used to interpolate meteorological variables for the HadUK-Grid dataset (Hollis et al., 2019).

Finally, similar to Stoner and Economou (2020a) and Oriani et al. (2018) the AME is implemented within a Bayesian hierarchical framework allowing for the full quantification of parametric and predictive uncertainty. As described by Oriani et al. (2018), Bayesian models produce posterior distributions of the model parameters, and hence sampling from these posterior distributions to generate ensembles of weather sequences robustly propagates model parameter uncertainty to the weather simulations. Furthermore, with a Bayesian hierarchical framework, expert knowledge about the weather system can be easily incorporated in the model structure and in the prior distributions (Oriani et al., 2018).

Figure 1 provides a summary of the methodological steps of the AME, used to firstly fit the framework and subsequently simulate from it. Firstly, a HMM is fit to daily rainfall data each location/site. These site specific models are then used to transform the daily rainfall at each site to standard Uniform margins using the probability integral transform, allowing for the relationship between all sites to be modelled using a multidimensional copula. To simulate from the AME framework, spatially coherent stochastic simulations are firstly generated from the fitted multidimensional copula, representative of the modelled sites. These simulations are then transformed to temporally coherent daily rainfall at each site using the site specific HMMs. Finally, these site simulations are interpolated to a high-resolution grid over the region of interest. The following subsections describe these methodological steps in more detail.

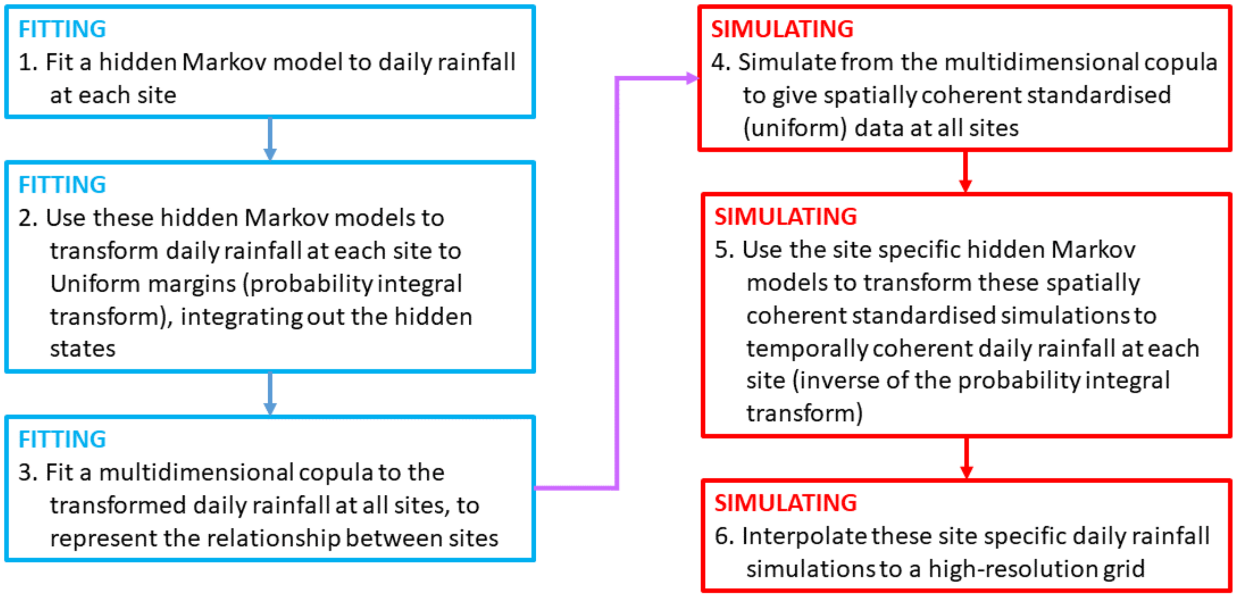


Figure 1. A diagram summarising the methodological steps of the AME, used to firstly fit the framework (blue boxes) and subsequently simulate from it (red boxes).

2.2 Modelling daily rainfall at each site: The hidden Markov model framework

Within the AME framework, the daily rainfall is modelled separately at each site using an extension of the advanced HMM proposed by Stoner and Economou (2020a). This is represented by step 1 in Figure 1. Capturing the whole distribution of rainfall (both low and high intensities) well can be challenging. This is particularly true when modelling rainfall at a high temporal resolution, such as on daily or hourly time-scales. An HMM aims to overcome this by representing the modelled variable (here rainfall) as coming from a discrete set of states (e.g. dry, wet and very wet). The variable is subsequently represented by a discrete mixture of distributions (rainfall models), each representing one of the hidden states. Following the notation of Stoner and Economou (2020a), the discrete random variable $z_t \in \{1,2,3\}$, representing the hidden state at time step t, is used to characterise the distribution of rainfall, x_t, at that time step as:

$$p(x_t) = \sum_{j=1}^{Z} 1(z_t = j) p(x_t|z_t), \tag{1}$$

where 1 is an identity function and $p(x_t|z_t)$ is the conditional distribution of rainfall x_t given the hidden state at time t, z_t. Temporal dependence is achieved within the HMM framework by assuming that z_t is an unobserved discrete Markov chain. This means that the temporal structure is captured by the persistence of each state, parametrised by the transition matrix $P = \{p_{i,j}\}$ where $p_{i,j} = \Pr(z_t = j|z_{t-1} = i)$, i.e. the probability of transitioning from state i into state j at time step t.

Stoner and Economou (2020a) show how a $Z = 3$ state HMM, characterising 'dry', 'wet' and 'wetter' conditions, is able to effectively model hourly rainfall behaviour in Exeter (UK). In addition, they show how including 'clone' dry states within the HMM framework provides additional flexibility. These multiple dry states are identical in their rainfall distribution model, but are allowed to differ in their persistence probabilities within the state transition matrix. Subsequently, the persistence distribution of the dry state has increased flexibility. For example, one of the clone dry states can capture shorter dry periods (i.e. have a lower probability of persisting), while the other can capture longer dry periods (i.e. have a higher probability of persisting).

In the case of a $Z = 3$ state HMM with one clone dry state, the resulting state persistence transition matrix can be summarised as follows:

$$P = \begin{pmatrix} p_1 & 0 & q_1(1-p_1) & q_2(1-p_1) \\ 0 & p_2 & q_1(1-p_2) & q_2(1-p_2) \\ v_1 r_{1,0} & v_2 r_{1,0} & r_{1,1} & r_{1,2} \\ v_1 r_{2,0} & v_2 r_{2,0} & r_{2,1} & r_{2,2} \end{pmatrix}, \quad (2)$$

where p_d is the probability of persisting in clone dry state d for $d = 1,2$; q_w is the probability of transitioning into wet state w, given a transition from dry to wet, for $w = 1,2$; $r_{i,j}$ is the probability of transitioning from state i to state j for $i, j = 0,1,2$ representing the dry, wet and wetter states respectively; and v_d is the probability of transitioning into clone dry state d, given a transition from wet to dry, for $d = 1,2$.

To ensure it is possible to interpret the clone dry states as a single state, a number of constraints are imposed on the state transition matrix. Firstly, the probability of transitioning between clone dry states is set to zero. Secondly, conditional on transitioning from dry to wet, the probabilities of transiting from the clone dry states into each of the wet states (q_1 and q_2) are the same for both clone dry states. Finally, conditional on transitioning from wet to dry, the probability of transiting from each of the wet states into the clone dry states (v_1 and v_2) are the same for both wet states.

Temporal variability in the transition matrix is captured through logistic regression on the dry state persistence probabilities, p_d for $d = 1,2$:

$$\log\left(\frac{p_d(t)}{1-p_d(t)}\right) = \iota(d) + f(x_1) + f(x_2) + \cdots, \tag{3}$$

where here, $\iota(d)$ is an intercept term, which is different for each clone dry state $d = 1,2$, and $f(x_i)$ for $i = 1,2,...$ are regression functions representing how temporally varying covariates $x_1, x_2, ...$ influence the dry state persistence probabilities. These covariates could be representative of temporal information such as ‘day of the year’, or climate drivers such as the NAO. These regression functions could take any number of forms, such as linear coefficients, regression splines or stochastic processes (e.g. GPs or random effects).

The conditional rainfall model in each hidden state consists of two parts: a model for the occurrence of rainfall; and a model for the intensity of rainfall (when it rains). As previously mentioned, the two clone dry states have the same conditional rainfall model.

The probability of zero rainfall at time t, denoted π_t, is modelled as varying in time, and is different for each state (e.g. dry, wet, wetter):

$$\log\left(\frac{\pi_t}{1-\pi_t}\right) = \eta(z_t) + f(x_1, z_t) + f(x_2, z_t) + \cdots, \tag{4}$$

where $\eta(z_t)$ is the intercept term, and $f(x_i, z_t)$ for $i = 1,2,...$ are regression functions representing how temporally varying covariates $x_1, x_2, ...$ influence the probability of zero rainfall in each state.

The intensity of rainfall in each state is modelled using a probability distribution. A number of different strictly positive distributions could be used here, such as the Gamma, Exponential or Generalised Pareto Distribution. These distributions are commonly characterised in terms of a scale and a shape parameter. The scale, σ_t, and shape, ξ_t, parameters are also modelled as varying in time, and are different for each state:

$$\log(\sigma_t) = \alpha(z_t) + f(x_1, z_t) + f(x_2, z_t) + \cdots \quad (5)$$

$$\xi_t = \gamma(z_t) + f(x_1, z_t) + f(x_2, z_t) + \cdots \quad (6)$$

where $\alpha(z_t)$ and $\gamma(z_t)$ are intercept terms, and $f(x_i, z_t)$ for $i = 1, 2, ...$ are regression functions representing how temporally varying covariates $x_1, x_2, ...$ influence the rainfall distribution parameters in state z_t.

The inclusion of independent effects for each state, and in all parameters of the rainfall model, results in a high degree of flexibility in the model. Further, applying this model within a Bayesian framework allows for the full quantification of parametric and predictive uncertainty. This also allows for additional model flexibility to be achieved through the implementation of Bayesian penalised regression.

2.3 Ensuring a transferable HMM structure: Bayesian penalised regression

Here, we extend the advanced HMM of Stoner and Economou (2020a) through the implementation of Bayesian penalised regression. This allows for a transferable model structure that can be applied without modification across many different sites/locations. Rather than requiring a model selection step prior to model fitting, to identify relevant covariates for explaining rainfall behaviour at each site, the penalised regression allows for model selection at the same time as model fitting. This additional flexibility is useful when modelling UK rainfall because the influential climate drivers of rainfall are known to vary across the country. For example, the NAO is known to have more influence in the west of the UK.

In a Bayesian framework, penalised regression is implemented through the prior distributions that are placed on the relevant regression coefficients/functions in Equations 3 - 6. These so-called 'shrinkage priors' in Bayesian penalisation aim to shrink small effects (i.e. unimportant covariates) to zero while maintaining true large effects (van Erp et al., 2019). Since these methods not only shrink the mean but also the variance of the posterior distribution of the unimportant effects to zero, their inclusion does not impact on the overall uncertainty of the model, and hence model performance.

Within the AME framework, the continuous horseshoe prior (Carvalho et al., 2010) is used to implement penalised regression, as it has been shown to be a good choice in terms of computational feasibility Piironen and Vehtari (2017). More technical detail about the continuous horseshoe Bayesian shrinkage prior has been provided in the Supplementary Material. The application of this within the AME framework to the Greater Anglian region is shown in Section 5.2.1.

2.4 Modelling the dependence between sites

When simulating alternative realisations of rainfall over a region, it is important that the model is able to capture the observed relationship between spatial locations. Within the AME framework, this is done using a copula approach (step 3 in Figure 1). Sklar's Theorem (Nelson, 2006) states that the 2-dimensional joint distribution of random variables x_1 and x_2 (here the rainfall at two sites), $F(h_1,h_2) = \Pr(x_1 \leq h_1, x_2 \leq h_2)$, can be written as:

$$F(h_1,h_2) = C\{F_1(h_1),F_2(h_2)\}, \tag{7}$$

where $F_1(h_1) = \Pr(x_1 \leq h_1)$ and $F_2(h_2) = \Pr(x_2 \leq h_2)$ are the univariate marginal distributions of x_1 and x_2 (here the HMMs fit to each site), and C is a copula: a 2-dimensional distribution on $[0,1]^2$ that represents the dependence between x_1 and x_2 when transformed to standard margins, achieved by applying the marginal distribution to each variable, also known as the probability integral transform.

Hence, in order to fit a copula model for the dependence between sites, the rainfall at each site must first be transformed to $[0,1]$ margins by applying the associated univariate marginal HMM distribution function, fit to that site (step 2 in Figure 1). The resulting transformed rainfall data are known as the 'residuals'.

When the univariate marginal HMM distribution function is applied to rainfall at a given site, the hidden state sequence must be integrated out in order to achieve the appropriate residual data for copula modelling. This is done by applying the marginal HMM distribution function conditioned on the most likely state sequence given the rainfall data, calculated using the forward-backward algorithm (Li and Jain, 2009). By integrating out the state sequence, however, the dependences between the hidden state sequences at different locations (i.e. the fact that being in the dry state at one site means the dry state is most likely in a neighbouring site) are not explicitly captured within the modelling framework. This is a necessary compromise of the approach because explicitly modelling the between-site state sequence dependence is computationally infeasible in a high-dimensional application (i.e when modelling more than a few sites together).

Stoner and Economou (2020b) proposed using a coupled HMM to explicitly model between-site state sequences for rainfall time series. Applied to daily rainfall time series from 3 sites in the UK, they were able to capture well correlations in occurrence and overall intensity, and joint exceedance probabilities between site pairs. However, this method involves constructing a combined state sequence with $(Z^S)^2$ states, where S is the number of modelled sites and Z is the number of modelled states at each site. Hence, explicitly modelling the between-site state sequence dependence at (for example) 30 different sites in a 4-state coupled HMM would require the estimation of $(4^{30})^2 = 1.3 \times 10^{36}$ state transition probabilities. This is computationally infeasible, hence representing the state sequences separately at each site, and modelling the dependence in the transformed rainfall residuals using a copula, is the computationally feasible alternative. In addition, it is expected that this state sequence dependence will be captured to some extent by fitting all 1-dimensional HMMs to spatially coherent rainfall data, and by regressing on the same temporal covariates in each of the transition matrix regression models (Equation 3).

Most often in spatially coherent rainfall models, such as Wilks (2008), Verdin et al. (2015), Kleiber et al. (2012) and Kroiz et al. (2020), the dependence between locations is modelled using a Multivariate Normal distribution (equivalent to a Gaussian copula), or a GP spatial dependence model. The Gaussian copula assumes asymptotic *independence* between random variables, meaning that the largest values of the variables are represented as rarely occurring together (Coles et al., 1999). The alternative, asymptotic *dependence*, may also occur in environmental data (Dawkins and Stephenson, 2018), meaning that the largest values of the variables tend to occur at the same time. For additional flexibility in the AME framework, different forms of dependence structure (asymptotic independence and dependence) between pairs of sites is captured. This flexibility is found to be important in the application of the AME framework to the Greater Anglian region of the UK (see Section 4.2). Within the AME framework, this is achieved by using a vine-copula dependence model (Czado, 2019).

Vine-copulas are a very flexible way of modelling dependence between multiple variables (in this case residual rainfall at different sites). Vine-copulas use a series of bivariate copulas within a nested set of 'trees' to model multiple combinations of pairs of variables (see Czado (2019) and Aas et al. (2009) for more detail). This flexibility is achieved within the vine-copula by allowing for each bivariate pair to be represented by a different form of copula dependence model. This means that, for example, the dependence between one pair of variables could be represented by the asymptotically independent Gaussian copula, while another pair could be represented by the asymptotically dependent Gumbel copula.

The AME framework employs a regular vine-copula (Dissmann et al., 2013). This form of vine-copula represents the dependence between N variables using $N-1$ nested trees. The first tree consists of $N-1$ bivariate models, linking all N variables together in a pairwise fashion. The second tree then consists of $N-2$ bivariate models, linking the $N-1$ pairs of variables together into pairs-of-pairs. This nested structure continues until the $(N-1)^{\text{th}}$ tree, which links together all N variables using a final bivariate model.

During model fitting, the way in which the pairs of variables are grouped (i.e. which are paired together in tree 1, tree 2 etc.) is determined by the maximum spanning tree (Monma et al., 1990), where the edge weights are equal to the empirical Kendall's tau (Czado, 2019). Within the AME framework, the vine-copula model fitting is given the option to represent the dependence between each pair (or pair of pairs etc.) in each tree using either a Gaussian or a Gumbel copula (other copula models could also be incorporated, such as the Frank or Joe copulas). During the vine-copula model fitting, the optimal copula model for each pairing (e.g. Gaussian or Gumbel) is determined by the Akaike information criterion (AIC) (Bozdogan, 1987). That is, each possible combination of copula is explored for each pair within each tree, and the optimal combination is chosen such that it best fits the data according to the AIC.

This fitted copula can then be used to simulate standardised (uniformly distributed) data spatially coherently across the modelled sites (step 4 in Figure 1). These copula simulations are then transformed to temporally coherent daily rainfall at each site by applying the associated site specific HMM quantile function (inverse cumulative distribution function), conditioned on the most likely state sequence and the historical covariates (step 5 in Figure 1).

It should be noted that, although the vine-copula provides a great deal of flexibility in the modelled dependence structure in a computationally efficient way, the vine structure is better suited to modelling tree-like organised data, such as flow along a river network (Timonina et al., 2015). Within the vine-copula model the cross-correlations between only N-1 pairs of variables are explicitly modelled, while the other hierarchical bivariate structures describe the links between probabilities resulting from the bivariate distributions at the lower stages of the hierarchy. When modelling rainfall, which might not necessarily have this tree-like structure, it may therefore be preferable to use copulas characterised by all pairwise cross-correlation matrices, such as meta-elliptical copulas (Serinaldi and Kilsby, 2014b). Meta-elliptical copulas are able to capture extremal dependence between variables; however, they require the specification of one asymptotic dependence structure for all pairs of variables. There is a growing literature in the area of flexible pairwise dependence models for extremal dependence (Wadsworth and Tawn, 2012; Huser and Wadsworth, 2019), though these models are non-trivial to apply and computationally challenging to fit to higher dimensional data (Dawkins and Stephenson, 2018). Their application is therefore found to be beyond the scope of this study. In addition, when applying the AME framework to the Greater Anglian region, the vine-copula was found to capture well the dependence in residual rainfall between sites not explicitly modelled as pairs within the vine structure. This supports the suitability of this flexible and computationally cheap alternative (see Section 4.2).

2.5 Interpolation

The desired final output of the AME framework is a high-resolution gridded daily rainfall dataset. To achieve this, the spatially and temporally coherent stochastic simulations at the modelled sites are interpolated to the desired grid (step 6 in Figure 1). This is done using the spatial interpolation method developed by the Met Office National Climate Information Centre (NCIC), previously implemented to interpolate meteorological variables for the HadUK-Grid dataset Hollis et al. (2019). This interpolation method is described in detail in Perry and Hollis (2005). For daily rainfall, the approach:

1. Calculates the long term average of daily rainfall at each station in the HadUK-Grid station record (Hollis et al., 2019);
2. Fits a regression model to these long-term averages using terrain elevation and terrain shape as covariates;
3. Interpolates the regression residuals to the target grid using inverse-distance weighted averaging;
4. Adds back the regression model covariates to give gridded long-term average rainfall;
5. For the site data to be interpolated (here the stochastic simulations at the 39 sites), calculates daily rainfall at each site and on each day as the percentage of the long-term average at that location;
6. Interpolates these rainfall percentages to the target grid using inverse-distance weighted averaging;
7. Multiplies the gridded long-term averages with the gridded rainfall percentages in each grid cell, to give gridded daily rainfall output;

Since the method utilises the long term average of the HadUK-Grid daily rainfall at each location across the region (point 1 above), the method retains the climatology (long term properties) of the original gridded data: a desirable feature of the approach.

Following this final step in the methodology (see Figure 1), the AME framework provides spatially and temporally coherent synthetic daily rainfall data on a high-resolution grid over a region of interest.

3 Data

The AME framework is now applied to a case study region in the east of the UK. This section introduces the data used within this application.

3.1 Rainfall data

Observed rainfall time series are taken from the HadUK-Grid daily rainfall dataset (Hollis et al., 2019), which covers the UK. This dataset is produced on a 1 km $\times$ 1 km grid resolution on the Ordnance Survey National Grid (OSGB) projection system, and we consider data in the period 1891-2018. From HadUK-Grid, a total of 39 daily rainfall time series for 39 sites across the Greater Anglian region are extracted. These 39 sites are distributed across the region (Figure 2) and cover a range of rainfall climatologies. The region covers some notably dry areas that receive little more than 500 mm of rainfall per year on average (represented by sites such as 25, 28, 32, 34 and 38). In contrast, some north-western areas of the region (representing the higher ground of the Pennines) see more than 1300 mm of rainfall per year on average (represented by sites such as 6 and 7). Further exacerbating the spatial variability in rainfall in this region is the coastline. There is more rainfall around north and east facing coasts, mostly due to convective rainfall that is triggered in the colder seasons over the North Sea. This makes for a complicated water resource situation. The majority of water supplies in this region come from aquifers and reservoirs. Groundwater sources are slow to refill. The nature of water resource here means that the region is particularly vulnerable to droughts arising from multi-month to multi-year rainfall accumulation deficits. This is particularly the case when there is a rainfall deficit in the winter months, since this is when most groundwater recharge occurs.

To ensure the 39 daily rainfall time series are of high quality, and can be taken forward in modelling daily rainfall at each site (Section 2.2), we conduct a number of quality checks. The 39 sites represent the locations of monthly station data available to Anglian Water and used in previous stochastic dataset generation. Therefore, aggregating the daily rainfall time series from HadUK-Grid to monthly time series should give similar values to those from the monthly station data. This should especially be the case where station data from the 39 sites is included in the HadUK-Grid daily rainfall dataset. If there are differences in a certain period, this could indicate that derived daily rainfall time series for the 39 sites rely more heavily on data interpolation, rather than observed station data. This is less suitable for modelling purposes since interpolated daily rainfall values are not likely to capture some of the highest rainfall totals that actually occurred, especially those associated with heavy convective showers.

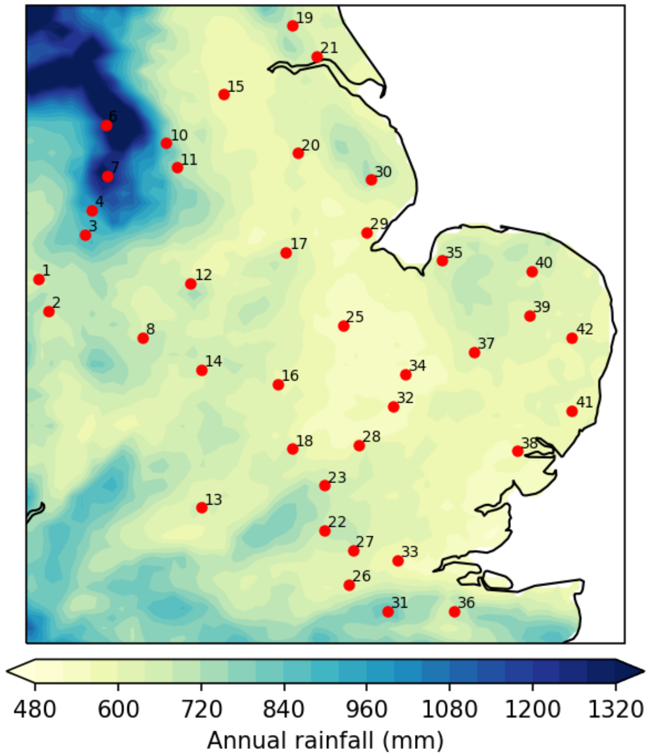


Figure 2. The location of the 39 sites across the Greater Anglian region, plotted on top of the annual mean rainfall (using a 1961-1990 climatology period). The 39 sites are not labelled sequentially, since they represent a subset of an original set of 42 sites. As such, the highest numbered site is 42.

The difference averaged across all 39 sites is relatively stable across the common time period of 1900-1990, although the differences appear noticeably larger before about 1910 (Supplementary Figure 1). Further data quality checks are undertaken to check for any unusual behaviour in the daily rainfall time series. This considers two key features of the daily rainfall distribution; the proportion of zero rainfall days and the variance of daily rainfall (calculated for each year). Both the proportion of zero rainfall days and the variance of daily rainfall are smallest in the early part of the record, noticeably before about 1915 (not shown). This is particularly stark for the variance of daily rainfall.

Although there is a possibility that these early 20th-century “trends” could be due to natural climate variability, several individual sites show a change in these two features of the daily rainfall distribution as a marked step change around 1910. This change in behaviour is not entirely unexpected; even though HadUK-Grid improves considerably on previous UK observed gridded datasets, in part through extensive digitisation of historical observational data, there are still inhomogeneities that can appear, most commonly through using a time-varying network of underlying rain gauges. Rainfall datasets are known to be particularly sensitive to changes in the distribution of rain gauges (Legg, 2015).

In HadUK-Grid the maximum number of rain gauges is reached in the 1960s and 1970s before numbers fall, albeit with a more optimal distribution of gauges (Hollis et al., 2019). Before this time gauge numbers are much reduced, especially those recording daily rainfall. Therefore, the HadUK-Grid gridded dataset relies more heavily on data interpolation before this time. An obvious outcome of this is to expect more homogeneous spatial information early in the 20th century, and this is evidenced by the reduced rainfall variability and proportion of zero rainfall days here. Given this information, and the wish to retain information from some high impact droughts in the Greater Anglian region in the 1920s and 1930s, this study makes use of the observed record in the period 1914-2018.

3.2 Covariates

For the calculation of climate indices/drivers, we use data from the European and North Atlantic daily to multi decadal climate variability (EMULATE) mean sea level pressure (EMSLP) dataset (Ansell et al., 2006) and the fifth generation of European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric reanalyses of the global climate (ERA5). EMSLP consists of mean sea-level pressure (MSLP) (as daily means) for the 1850–2003 period covering the North Atlantic and European (NAE) region (25°–70°N, 70°W–50°E). ERA5 consists of MSLP (as daily means) for the 1979–present period with global coverage. This allows for continuous MSLP indices to be calculated for the period 1914-2018, using ERA5 for the period 2004-2018. There is strong agreement between the two MSLP datasets in the common period, 1979–2003 (not shown). For calculating an SST index we use the Hadley Centre Sea Ice and Sea Surface Temperature (HadISST) dataset (Rayner et al., 2003), over the period 1914-2018.

Including the most relevant explanatory variables (covariates) to represent the drivers of rainfall in the Greater Anglian region is crucial for simulating realistic rainfall variability and, in turn, hydrological impacts, such as long duration (multi-year) UK drought. For example, Serinaldi and Kilsby (2012) used the NAO and SSTs as explanatory variables for winter monthly rainfall. These were shown to influence rainfall variability across the UK. However, because the focus in this work is on one specific region of the UK, the Greater Anglian region, this topic is revisited to understand whether there are stronger drivers of rainfall variability here. The need for this was highlighted by Lister et al. (2018), suggesting that using stronger predictor-precipitation relationships could further improve the performance of rainfall generators. The inclusion of the EA pattern in configurations of rainfall generators has improved performance in previous industry-led work undertaken by the Met Office in the Greater Anglian region, as detailed in Lister et al. (2018). For this reason the EA pattern was also considered as a predictor in this study.

Recent research has added more clarity on these relationships and how the inclusion of explanatory variables could be optimised. For example, the NAO is a useful explanatory variable in winter but the summer North Atlantic Oscillation (SNAO) – the summer counterpart of the NAO – still explains 25% of northern European summer rainfall variability (Dunstone et al., 2018). A more favourable representation of this dominant mode of summer variability could lead to improved simulation of temporal persistence of below average rainfall across the summer months, representative of realistic shifts in the jet stream and prolonged periods of blocking regimes. The summer of 2018 over the UK and northern Europe is one such example of a strongly positive SNAO season, which was exemplified by an amplified and persistent omega block (Kornhuber et al., 2019). Where NAO is used as an explanatory variable year-round, it often uses a station-based NAO index (e.g. the difference between the normalised MSLP over the Azores and the normalised MSLP over Iceland). This definition is trained on the winter NAO and will not be as strongly correlated with summertime UK temperature and precipitation (Hall and Hanna, 2018). There is a stronger correlation between the SNAO and precipitation in southern and eastern areas of the UK, relative to northern and western areas.

One objective of the AME development is to pick explanatory variables that show strong relationships with precipitation variability across many parts of the UK throughout the year, so that the AME approach is suitable for application across all UK catchments and water regions. The EA pattern and NAO offer different levels of precipitation correlation in different regions of the UK. This correlation also varies by season. Therefore, a more robust representation of these circulation indices in stochastic weather generators is important. Here, four relevant climate indices are explored.

3.2.1 MSLP-based indices

Three indices are derived from MSLP and represent key modes of pressure variability in the North Atlantic-European (NAE) domain; the winter NAO (WNAO), the summer NAO (SNAO) and the EA pattern. Empirical orthogonal function (EOF)/principal component (PC) analysis is used to understand and calculate the three MSLP-based indices. EOF1 for winter shows the well-documented WNAO pattern, with pressure dipoles to the west of Spain and near Iceland (Figure 3). This particular example (Figure 3a) shows a positive WNAO pattern, which would be characterised by warmer, wetter and windier conditions over the UK. EOF2 for winter (Figure 3c) shows the EA pattern, with anomalous pressure over a broad region of the NAE domain, centred between the UK and Iceland; this shows the EA pattern in its positive phase, characterised by more unsettled weather over the UK. In contrast, EOF1 for summer shows the SNAO (Figure 3b). The SNAO is clearly shifted northwards relative to its winter counterpart, with centres of action over the UK and Greenland. Here, the positive phase of the SNAO is shown, relating to warmer, drier conditions for the UK. EOF2 for summer still pertains to the EA pattern (Figure 3d). It retains a centre of action between the UK and Iceland, albeit shifted slightly northwards.

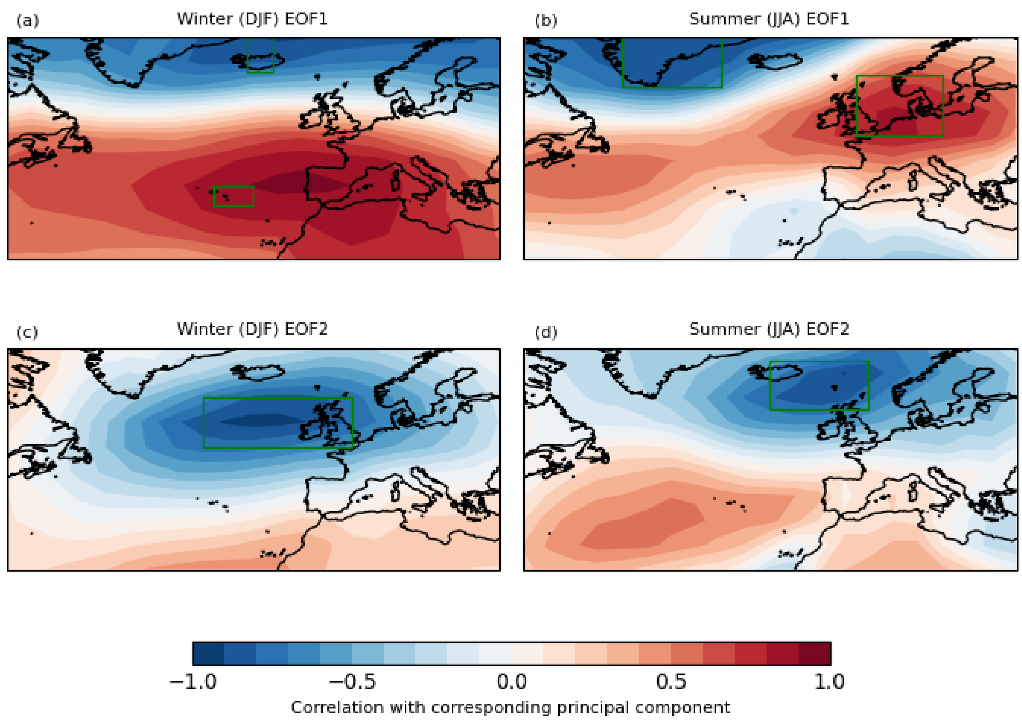


Figure 3. EOF1 in (a) winter (December-January-February), representing the WNAO and (b) summer (June-July-August), representing the SNAO, and EOF2 in (c) winter and (d) summer, both representing the EA pattern. The green boxes show the domains over which MSLP is calculated in deriving the indices.

A traditional approach to calculating the MSLP-based indices is used, calculating either the average MSLP over a domain (EA pattern), or the MSLP difference between two domains (SNAO and WNAO indices). Results are similar if using an EOF-based approach and projecting spatial patterns of daily MSLP anomalies on the EOFs of the circulation patterns. The domains used to calculate indices, overlaid on the EOFs, are shown in Fig. 3. The WNAO and SNAO indices are calculated as the difference between the southern node and the northern node (the WNAO reflects the common definition of taking the pressure difference between the Azores and Iceland). Indices are calculated at daily frequency. The WNAO is calculated for all days in the winter half-year (October-March) and the SNAO is calculated for all days in the summer half-year (April-September). The EA pattern is calculated as a continuous index covering all seasons, although we use slightly different definitions in the summer-half year and the winter half-year, reflecting the slight shift in the centre of action of this mode of variability between summer and winter (Fig. 3c,d).

3.2.2 SST index

Recent work has shown that there is a relationship, albeit a weak one, between North Atlantic SSTs and rainfall over the UK, as well as a wider Northern European area (Sutton and Dong, 2012; Dunstone et al., 2018). Anomalous North Atlantic SST states have been proposed as a mechanism for driving long-duration rainfall deficits and meteorological drought (Sutton and Hodson, 2005). Here, we determine a domain to use for calculating an SST index by regressing monthly North Atlantic SSTs (following linear detrending, to remove any signal due to global mean warming) against a Greater Anglian regional average monthly rainfall time series.

Considering all months (rather than individual seasons), there is a weak but significant relationship between rainfall variability in the Greater Anglian region and SST anomalies over a region centred to the south-west of the UK (Figure 4). This same pattern emerges when regressing against both the proportion of zero rainfall days and the variance of daily rainfall. These SST anomalies are even more pronounced during the peak long-duration (one year or longer) drought events experienced in the 1914-2018 period over the Greater Anglian region, with these SST anomalies mirrored when looking at long-duration wet periods (not shown). Given this region shows the strongest correlation with rainfall variability over the Greater Anglian region, a monthly SST index is calculated as the average anomalies over the domain 45.5°-54.5°N, 20.0°-5.0°W.

4 Model set up

This section describes how the AME framework is applied to model and simulate daily rainfall across the Greater Anglian region.

4.1 Modelling daily rainfall at each site

Daily rainfall data is modelled at each of the 39 sites presented in Figure 2, using the HMM framework described in Section 2.2. Specifically, a $Z = 3$ state HMM was used here, as it has previously been shown to represent hourly rainfall behaviour in Exeter (UK) well (Stoner and Economou, 2020a). To minimise model complexity we began by including just one additional clone dry

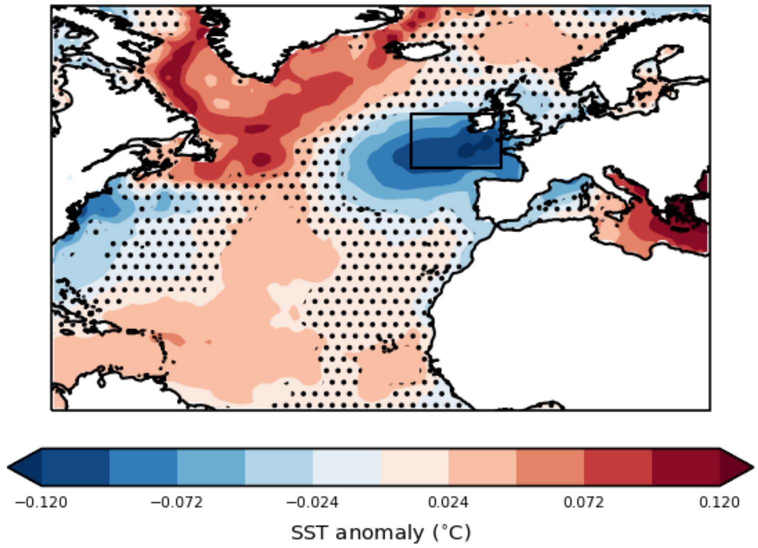


Figure 4. Regression of monthly SST anomalies onto the Greater Anglian region monthly rainfall time series. Units are per standard deviation of the Greater Anglian regional average monthly rainfall time series. Stippling indicates where the regression is not statistically significant at the 95% confidence level and the black box represents the domain over which the SST index is subsequently calculated (45.5°-54.5°N, 20.0°-5.0°W).

state (rather than the two used in Stoner and Economou (2020a)). Following model checking, it was concluded that this number of clone dry states was adequate for capturing dry periods in the observations. In addition, it might be expected that less dry state persistence flexibility would be required to model *daily* rainfall here, compared to *hourly* rainfall in Stoner and Economou (2020a), hence the need for fewer clone dry states.

The mathematical configuration of equations 3 - 6, developed for this case study application, are given in the Supplementary Material. The temporal variability in the HMM state transition matrix, captured through the logistic regression on the dry state persistence probabilities (equation 3), is modelled as a function of ‘time of the year’ and ‘time overall’. Specifically, a cyclic cubic spline of the ‘time of the year’, capturing the smooth seasonal variation in the persistence of the dry states, and an independent and identically distributed (i.i.d.) stochastic process of 5-yearly Gaussian random effects, capturing the non-smooth, between-year variability in the persistence of the dry states. Each 5-yearly random effect represents one 5-year block of data in the period 1914-2018, and the same random effect value is representative of all days in that 5-year block. An alternative model using 1-yearly random effects (i.e. a single value representative of all days in each year) was also explored, but was found to provide no additional benefit in representing rainfall behaviour. The 5-yearly random effects model is therefore used to reduce model complexity. Using random effects rather than smooth regression splines to capture this ‘time overall’ effect also improves model robustness. This is because the random effects are able to successfully capture the observed non-smooth 'jumps' in dry state persistence in time (as seen in Figure 8 in Section 5.2.1).

The temporal variability in the probability of zero rainfall in each hidden Markov state (equation 4) is modelled as a function of 'time of the year' and the four climate drivers discussed in Section 3.2. As above, the 'time of the year' effect is captured using a cyclic cubic spline, while the relationship between each climate driver and the (logit) probability of zero rainfall in each state (equation 4) is captured by a single regression coefficient, characterising a linear relationship.

As described in Section 2.2, the intensity of rainfall in each state is modelled using a probability distribution. In line with the model developed by Stoner and Economou (2020a), and in order to capture extremes well, we model rainfall in each state using a zero-threshold Generalised Pareto Distribution (GPD). The scale and shape parameters of these distributions, representing the variance and tail behaviour respectively, are modelled as varying in time (equations 5 and 6). The scale parameter is modelled as a function of 'time of the year', 'time overall' and the four climate drivers, while the shape parameter is modelled as a function of 'time of the year' only. As above, in both cases the 'time of the year' effect is captured using cyclic cubic splines, the 'time overall' effect is captured using i.i.d. 5-yearly Gaussian random effects, and the climate driver effects are represented by linear regression coefficients. Fewer effects are included in the model for the GPD shape parameter because this parameter is known to be more difficult to estimate. Including fewer effects here therefore improves the robustness of model fitting.

The combination of temporal effects, characterised by seasonal splines and 5-yearly random effects, were chosen for each part of the model following a model selection process. This was focused on a subset of the 39 sites, representing a range of rainfall climatologies. The temporal effects were included one-by-one and the ability for the resulting model to simulate rainfall with accurate temporal variability in the proportion of zero rainfall days in each year and the rainfall variance in each year (compared to observations) was explored. The combination of temporal effects that minimised model complexity but adequately represented these rainfall characteristics was chosen as optimal.

There is a wealth of literature on the relationship between UK rainfall and large scale atmospheric circulation patterns, consisting of established theory and emerging ideas (Wilby et al., 1997; Folland et al., 2009; Ossó et al., 2018). Within this case study application of the AME, we include the effect of four climate drivers: WNAO, SNAO, EA pattern and SSTs, tailored to the region of interest, as described in Section 3.2. The climate drivers are each scaled by their mean and standard deviation to ensure they are on a consistent scale. A seasonal indicator variable is used to ensure that SNAO/WNAO is only included within the regression model in the summer/winter. Here, summer is defined as the 6-month period April-September (inclusive) and winter is defined as October-March. It was found that rainfall simulations from configurations of the model not including these climate drivers were less able to reproduce the most extreme meteorological drought events in the historical record. This indicates the added value of including these effects within the model.

When fitting this HMM to each site in the region of interest, the same temporal effects are included in equations 3, 4, 5 and 6, as it is expected that seasonal and annual variability in rainfall behaviour will be experienced throughout the region. The influential climate drivers of rainfall are, however, known to vary across the UK. For example, in the most easterly part of the Greater Anglian study region, the EA pattern is known to be important for representing rainfall conditions, while the WNAO has little influence. In the most westerly part of the region, however, WNAO is known to be a more important driver of rainfall behaviour. The Bayesian penalised regression method described in Section 2.3 is used to select which of these climate drivers is relevant in each location/site model during model fitting. This approach allows the model to 'shrink' away the effect of unimportant climate driver covariates leaving just those that are important at that location. This provides a generic model structure that can be fit to all sites (both in the East and West of the Greater Anglian region), but that allows for any combination of the four climate drivers to be included in the HMM at each site. The results of this penalised regression are presented in Section 5.2.1.

This configuration of the HMM is fit to the 105 years (01/01/1914-31/12/2018) of daily rainfall data (as introduced in Section 3), separately for each of the 39 sites in the Greater Anglian region. This is done within a Bayesian statistical modelling framework. As such, prior distributions must be specified for each model parameter. The priors used in this application are presented and discussed in detail in the Supplementary Material.

To ensure the desired three state structure (dry, wet, wetter), and avoid hidden state label switching (Jasra et al., 2005) during model fitting, a number of constraints are imposed upon the model parameters (as defined in equations 3 - 6) during model fitting:

$$\iota(1) > \iota(2); \tag{1}$$

$$\eta(\text{dry}) > \max(\eta(\text{wet}), \eta(\text{wetter})); \tag{2}$$

$$\eta(\text{wet}), \eta(\text{wetter}) < 0. \tag{3}$$

$$\alpha(\text{wetter}) > \alpha(\text{wet}); \tag{4}$$

$$\gamma(\text{wet}) < 0.25, \tag{5}$$

where $\iota(d)$ is the intercept term of the logistic regression model for the dry state persistence probability for clone dry state $d = 1, 2$; and $\eta(z_t)$, $\alpha(z_t)$, and $\gamma(z_t)$ are intercept terms for the regression models for the probability of zero rainfall, the rainfall distribution scale parameter and the rainfall distribution shape parameter (for hidden state z_t: dry, wet or wetter) respectively. The first four constraints enforce that, on average (since they are applied to the regression intercept terms), clone dry state 1 should be more persistent than clone dry state 2 (constraint 1); the probability of zero rainfall should be higher in the dry state than in the two wet states (constraint 2); it should be more likely to rain than not rain in the two wet states (constraint 3); and the rainfall variance is higher in the wetter state than the wet state (constraint 4).

As well as these four state identifiability constraints, an additional constraint is placed on the shape parameter of the 'wet' state rainfall distribution (constraint 5). This constraint is included to limit the upper tail of the GPD distribution. Due to the heavy-tailed nature of the empirical rainfall distribution (meaning that the distribution/histogram of the data goes to zero slower than the exponential distribution), the GPD shape parameters is predominantly estimated to be positive in order to capture the few higher daily rainfall amounts. When the GPD shape parameter is positive the associated GPD distribution has an infinite upper tail. This means that, although very unlikely, an infinitely high rainfall value can be simulated. When fitting this model and simulating daily rainfall from it within the Greater Anglian region, this was found to lead to (very infrequent) simulation of extremely high daily rainfall values, in excess of 1000 mm in one day. To overcome this, a realistic upper limit for daily rainfall within the region was explored, and used as a guide to tune the Bayesian model structure, leading to the specification of constraint (5). This constraint is consistent with the values resulting from the global multiple threshold method analysis reported in Serinaldi and Kilsby (2014a).

Taking a UK-wide historical context the highest 24-hour rainfall total in the UK was 341.4 mm commencing at 18:00 GMT on 4 December 2015 at Honister Pass, Cumbria. The highest amount falling in a 'rainfall day' (0900-0900 GMT) in the UK is 279 mm, on 18 July 1955 in Martinstown, Dorset. In the Greater Anglian region, the 25-27 August 1912 extreme rainfall event is the largest observed storm, where over 200 mm was recorded (Brooks, 2012). It is difficult to argue that the Honister Pass rainfall could have occurred in the Greater Anglian region, due to the lack of orography here, which was an important factor in enhancing the rainfall during the Honister Pass event. However, one could argue that a similar synoptic scenario to that observed in the Martinstown storm could also occur across the Greater Anglian region. This agrees well with extreme value studies and catalogues for the Greater Anglian region, giving long return period 1 in 10,000 year event estimates for daily extreme rainfall of around 300 mm (Francis, 2011). It was found that, for some locations, the daily rainfall simulated from the HMM without constraint (5) was very occasionally (approximately 0.01% of simulations) higher than this value. These very high simulations were found to all occur in the 'wet' state and when the wet state shape parameter was greater than 0.25. Imposing constraint (5) was subsequently found to greatly reduce the number of simulations in excess of this threshold. It should be noted, therefore that this 300 mm threshold was not used as a truncation, but as a guide for tuning the Bayesian model structure. Rainfall simulations in excess of this threshold are still feasible within the model, as is the case in the real world given that 300 mm does not represent a physical barrier. Indeed, rainfall maximum thresholds should be used with care to ensure environmental risk is not under- or over-estimated (Yevjevich, 1968).

For each model fit, four MCMC chains were run in parallel, each for 30,000 iterations. Model exploration identified that this number of iterations were required to reach convergence. In each case the first 20,000 iterations were discarded as burn-in, and the remaining 10,000 were thinned by 40 (i.e. taking one sample in every 40), leaving 1000 samples in total from the four chains, representing 1000 samples from the posterior distribution of each model parameter. These 1000 samples are used to quantify parameter uncertainty in the model checking and validation (see Section 5).

Due to the complexity of the model and the length of the data record being modelled (38,351 days), each of the 39 site models took approximately 40 hours to fit on a high-performance computer cluster using 8 Central Processing Units (CPUs) and 50 Gigabytes of memory. Convergence of the four MCMC chains was assessed by visual inspection of trace plots and by computing the Potential Scale Reduction Factor (Brooks and Gelman, 1998).

4.2 Modelling the dependence between sites

Fitting a regular vine-copula model to the residual rainfall at the 39 sites in our study region, as described in Section 2.4, results in a 38-tree nested structure. This vine-copula model takes approximately 8 minutes to fit to the full 105 years of daily data at all sites. A majority of the time a Gaussian copula is selected, particularly in trees 2-38, with the Gumbel copula mostly being selected to represent sites in close proximity, represented by tree 1.

Figure 5 shows the first two nested trees in the 38 tree regular vine-copula structure fitted to the daily residual rainfall at all of the 39 sites (note, the numbers on the tree are not the site numbers, how the tree values map to the site numbers is shown on the left of the plot). In tree 1, the 39 sites are linked together by 38 bivariate copula models. In all cases the dependence between each pair is modelled using a Gumbel copula. This is likely to be because all of these initial pairings are sites that are close together in location (and hence have stronger dependence in extreme rainfall, as seen in Figure 6 (a)). In tree 2, these 38 pairs of sites are then paired with another pair of sites. The dependence between these pairs of pairs are all modelled using a Gaussian copula, capturing how the dependence structure is different for sites that are less close together. Although trees 1 and 2 use the same copula (Gumbel and Gaussian respectively) for all pairings within, this is not always the case; predominantly pairings in the higher order trees use the Gaussian copula, but there are cases where one tree uses both Gumbel and Gaussian copulas for different pairs.

This vine-copula model is able to flexibly and accurately capture the differing asymptotic dependence structures between different pairs of sites in the region, using a series of bivariate copula models. This can be seen in Figure 6, where the relationship between simulated residuals at sites V38 and V41 (close together) capture the asymptotically dependent (Gumbel) dependence structure seen in the observed residuals (Figure 6 (a)), while the relationship between simulated residuals at sites V38 and V7 (far apart) capture the more elliptical asymptotically independent (Gaussian) dependence structure seen in the observed residuals (Figure 6 (b)). This highlights how using a model that assumes a Gaussian dependence structure between all pairs of locations would not be suitable in this application.

An additional plot is presented in the Supplementary Material (Supplementary Figure 2), comparing how the Gaussian copula (Multivariate Normal distribution) and the vine-copula capture the dependence in residual rainfall for site V38 paired with all other sites. This shows that for sites in close proximity to V38, the Gaussian dependence structure is inadequate, further

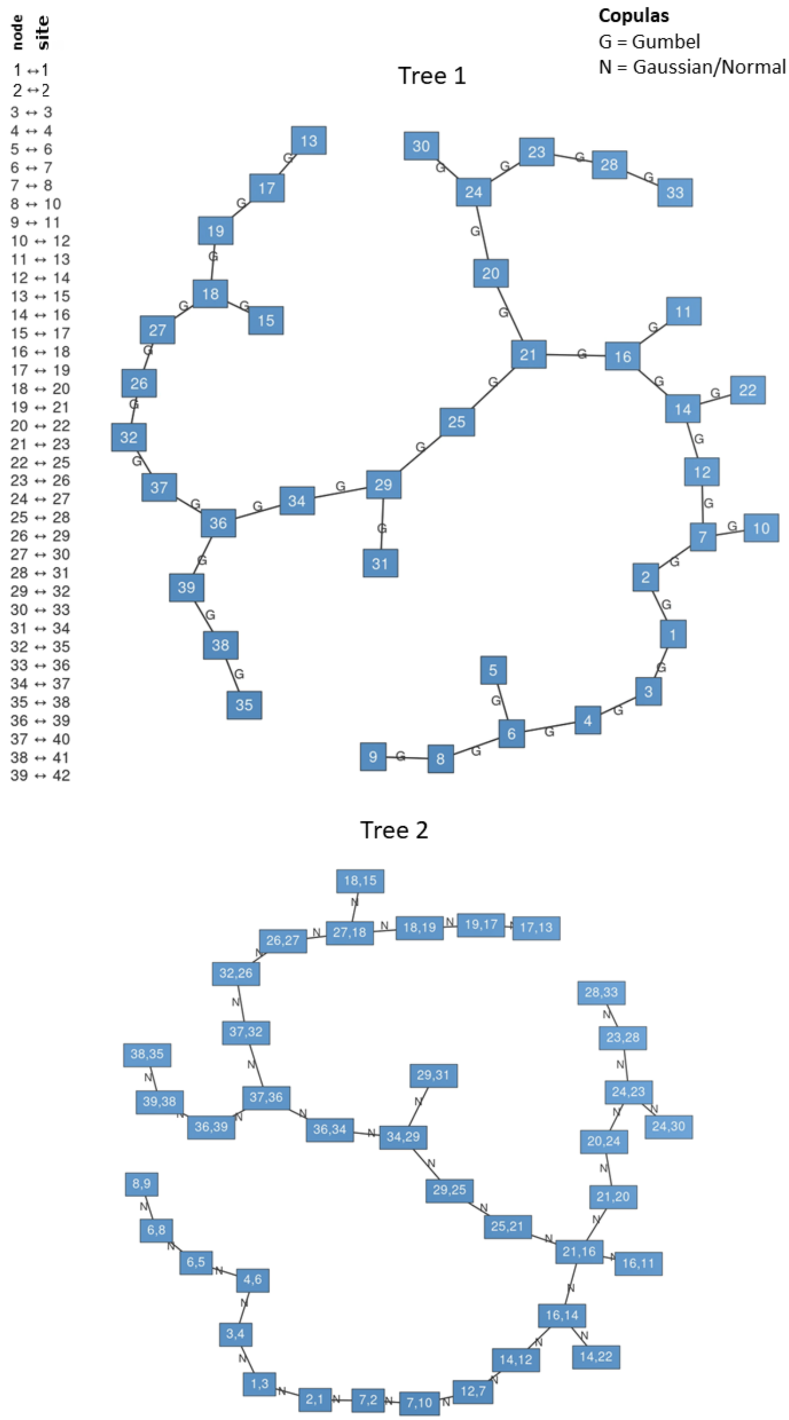


Figure 5. A graphical representation of the first two nested trees in the 38 tree vine-copula structure fitted to daily rainfall at all of the 39 sites transformed to standard Uniform residuals using the HMM at each site and one posterior parameter sample (chosen at random.) The nodes on the trees are numbered 1-39, and their associated site number (as shown in Figure 2) is presented in the legend on the left side. Each bivariate copula model in each tree is represented by an edge linking together two nodes. Each edge is labelled by the copula model used to represent that pairwise dependence, either the Gumbel copula (G) or the Gaussian/Normal copula (N).

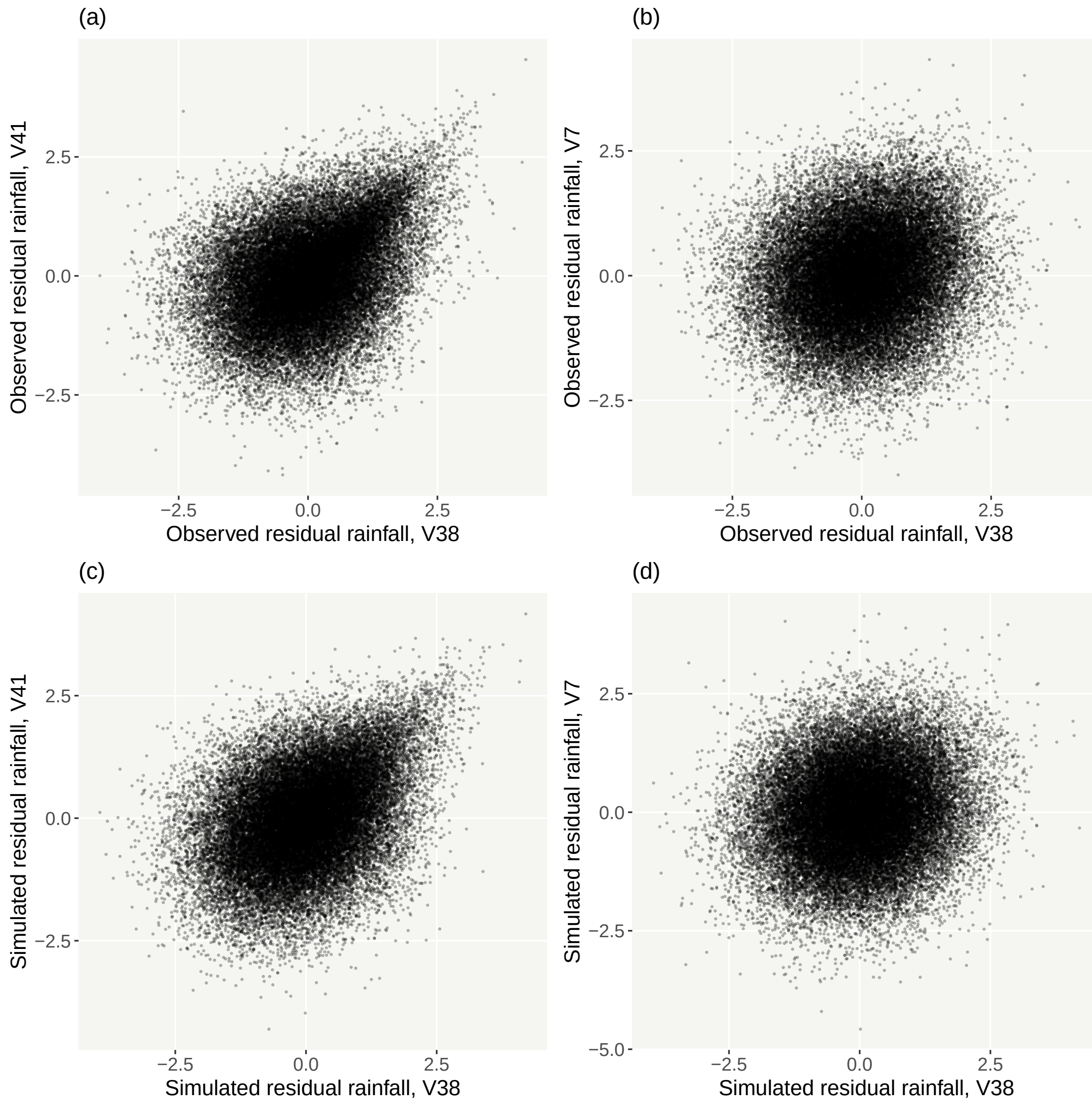


Figure 6. Top row: scatter plots showing the relationship between observed residual rainfall transformed to standard Normal margins (using one randomly chosen posterior sample of the HMM parameters) at pairs of sites within our study region (a) close together in space (approximately 30 km apart), and (b) far apart (approximately 240 km apart). Bottom row: equivalent for residual rainfall simulated from a vine-copula dependence model, fitted to all 39 sites, for the same pairs of sites (c) close together, and (d) far apart. In both cases, the close together sites are represented by V38 (Ipswich) and V41 (Woodbridge), and the far apart sites are V38 and V7 (Buxton).

supporting the use of the flexible vine-copula. Further, this additional figure and Figure 6 (b)/(d), show how the pairs of sites that are not explicitly modelled as pairs within the vine structure are still well represented by the vine-copula.

4.3 Simulation and interpolation

A single synthetic simulation of the 105-year historical period (1914-2018) is generated for each of the 1000 Bayesian posterior parameter samples from each site HMM, capturing stochastic and parametric uncertainty (see Section 4.1). This is achieved in a spatially coherent way across sites by following the steps in Figure 1, described in more detail in the Supplementary

Material. This approach of generating a single realisation for each MCMC parameter set is theoretically justified by Monte Carlo integration as a general approach for computing any posterior predictive quantify in Bayesian models, as described in chapter 1.9 of Gelman et al. (2014). This approach performs numerical integration to "integrate out" sampling uncertainty when producing predictions and is common practise in Bayesian modelling and Bayesian weather generators (see for example Section 4.3 of Verdin et al. 2019).

In each case, the same temporal (historical) pattern of climate drivers is used. This achieves good agreement between meteorological drought behaviour in the historical period and the simulations (as will be shown in the model validation in Section 5). An alternative pattern of climate drivers could, however, be used as discussed in Section 6. The subsequent dataset consists of 1000 alternative realisation of the 105-year period of daily rainfall, stochastically simulated spatially and temporally coherently at the 39 sites.

The interpolation method described in Section 2.5 is applied to each of the 1000 simulations and 38,351 days in the 105-year period separately. The simulated rainfall at the 39 sites is interpolated to the 57 $\times$ 52 grid shown in the Supplementary Material (Supplementary Figure 3), with latitude extents of 51.50-53.70°N and longitude extents of -2.39°W-1.76°E, where each grid box is 5 $\times$ 5 km in size. The 5 km grid is on the OSGB projection, the projection that is used in the NCIC interpolation scheme. These latitude and longitude extents transform to the nearest grid cell centres in northings of 177500-432500 and in eastings of 372500-652500. Finally, any grid cells that contain no land area are masked out as sea.

5 Model Checking and Validation

This section illustrates the performance of the AME framework, when applied to the Greater Anglian region, through a series of model checking and validation metrics.

5.1 Metrics

Model checking is based on the exploration of a selection of HMM temporal effects and estimated parameters:

1. The effect of 'time of year' on rainfall behaviour at site V38 (Ipswich). This site is used as an example because it is known to be one of the driest in the Greater Anglian region, and hence important to capture well when modelling drought;

2. The effect of 'time overall' on rainfall behaviour at site V38 (Ipswich);

3. The Bayesian penalised regression results, showing which of the climate drivers are included/excluded from each site HMM.

The spatially coherent rainfall simulations at the modelled sites, produced at step 5 of the methodology shown in Figure 1, are then compared with the observed rainfall record to validate how well the simulations capture key rainfall behaviour. This is done in terms of:

4. Quantile-quantile plots of daily simulated and observed rainfall in each season;
5. Quantile-quantile plots of monthly and annual total simulated and observed rainfall;
6. A comparison of the proportion of zero rainfall days in the simulated and observed records in each season;
7. Quantile-quantile plots the number of consecutive days with zero rainfall in the simulated and observed records;
8. A comparison of the 36-month accumulated rainfall Deficit Drought Index (DDI) in the simulated and observed records, presented as a time series and a quantile-quantile plot. This index is calculated by accumulating rainfall (observed or simulated) over each 36-month moving window within the 105-year period, and scaling this accumulated series by its long-term average and standard deviation (Burke et al., 2010);
9. Comparison scatter plots of monthly total simulated and observed rainfall at pairs of locations;
10. A comparison of the observed and simulated pairwise cross-correlation in daily (non-zero), monthly total and annual total rainfall as a function of separation distance.

For metrics 4-9, these plots are shown for four sites: V7 (Buxton), V16 (Ruthamford), V20 (Lincolnshire) and V38 (Ipswich). These sites are chosen to represent a range of rainfall climatologies within the regions (e.g. the Buxton site is characteristic of the wetter part of the study region in the west, while the Ipswich site is one of the driest). Equivalent plots for all 39 sites are available in the supplementary material (Supplementary Figures 6-12).

Finally, the gridded simulations, produced at step 6 of the methodology shown in Figure 1, are validated. This is done in terms of:

11. A comparison of the 36-month accumulated rainfall DDI in the simulated and observed records for the Greater Anglian region as a whole;

12. A comparison of the 36-month accumulated rainfall DDI, and subsequent drought duration, magnitude and severity, in the simulated and observed records for six independent rainfall sites (not included within the model fitting), with an emphasis on sub-regions that are not well represented by the 39 sites;

13. A qualitative assessment of the simulation performance in dry/wet periods and summer/winter seasons based on animations of rainfall maps.

To carry out the validation at the six independent rainfall sites, site data was retrieved from the Met Office Integrated Data Archive System (MIDAS) dataset, consisting of quality controlled rain gauge sites that form part of the UK rain gauge network, including official Met Office observation sites. The MIDAS data is used as an independent check on the simulated, gridded data that was produced using data extracted for 39 sites across the Greater Anglian region from HadUK-Grid (Section 3). While a number of checks can be undertaken, here the emphasis is again on the ability to simulate observed meteorological drought characteristics.

A number of regions were carefully selected, through an understanding of where there might be weaker performance of the simulated gridded output, namely where there are sub-regions not well covered by the existing 39 sites. This could be either because of gaps in the coverage or because they lie beyond the extents of the coverage; for example, coastal areas. Gaps exist because good quality, long records were not available in these areas. The location of the six independent rainfall sites is shown in Supplementary Figure 14. They cover three areas that were targeted for validation, in order of priority:

- Essex (in the southeast of the Greater Anglian region), which is not well represented by the 39 sites. This is arguably the most important area for validation, since it is one of the driest in both the Greater Anglian region and the UK as a whole. The sites Writtle and Shoeburyness are used for this validation. Shoeburyness also serves as a validation at a coastal site (see next point);

- Coastal sites, since these are not represented by the 39 sites. The expectation is that rainfall in coastal areas is not as robustly simulated as, for example, inland areas in close proximity to (and surrounded by) a number of sites. Cromer is used for this validation;

- Areas across Suffolk and Cambridgeshire that are not as well represented by the 39 sites. The sites Wattisham, Brooms Barn and Monks Wood are used for this validation;

Long, daily and digitised rainfall records are difficult to find in the MIDAS dataset. Most digitised records start in the 1960s and this is the case here, with the exception of Shoeburyness, for which extracted data starts in 1931. The sites were chosen as those that fell broadly in the sub-regions listed above, while having a record beginning in the 1960s or earlier and having nearly complete records. Where daily measurements were missing, those days were in-filled using the measurements from a nearby station.

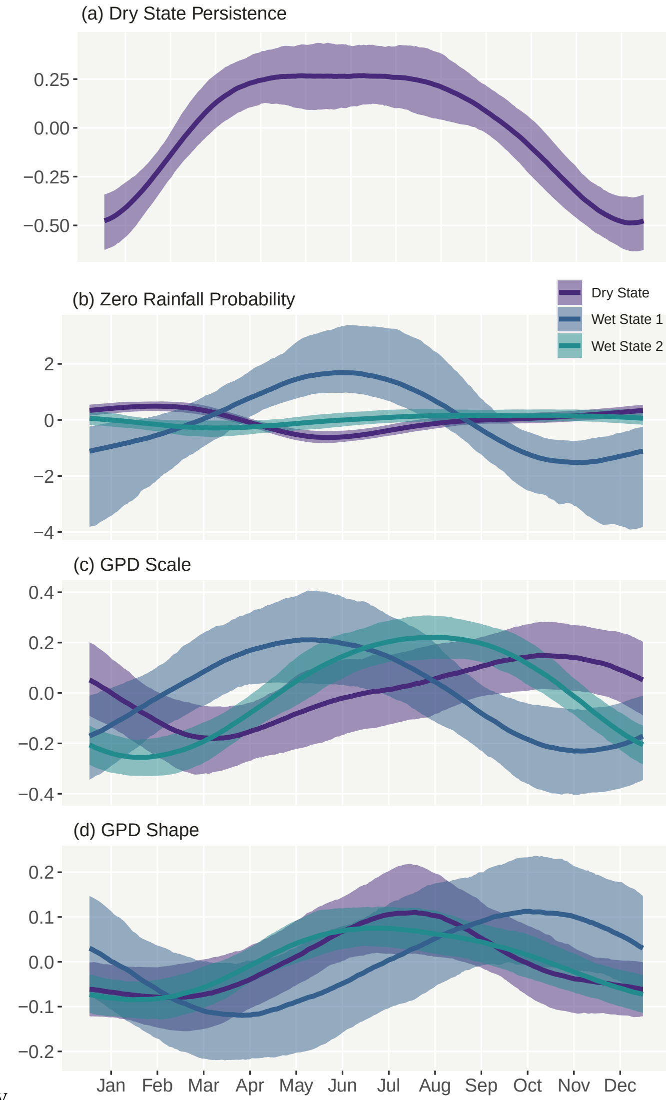
Rather than using the standardised DDI, a fairer comparison in this part of the validation is likely to be achieved through use of the DDI without standardisation, reflecting absolute deficits (and surpluses). This is to ensure that the full month-to-month rainfall variability is represented at these sites and they do not simply tend to climatology (the interpolation scheme uses long-term average rainfall as a starting point, as detailed in Section 2.5). If they were to tend towards climatology with dampened variability, the absolute deficits would be underestimated.

5.2 Results

5.2.1 Model Checking

Figures 7 and 8 present the effect of 'time of year' and 'time overall' respectively, on rainfall behaviour at site V38 (Ipswich). Figure 7 (a) shows how the model captures the expected seasonal variation in the persistence of the dry state, with this persistence being higher in the drier late-spring and summer months. This trend is also seen in the other 38 site HMMs (not shown).

The rainfall distribution parameters, shown in Figure 7 (b)-(d), vary throughout the year in a different way in each state. In summer, in the dry state, the zero rainfall probability is slightly lower, combined with a lower rainfall variance (scale parameter) and higher rainfall distribution tail (shape) parameter. This suggests that, due to the increased persistence and hence more frequent occurrence of the dry state in the summer, summer-time rainfall occurs within this state, and that this rainfall is either relatively low (low scale parameter) or extreme (high shape parameter). This could be linked to the known UK summer-time rainfall behaviour associated with convective storms which bring short heavy downpours in summer. Interestingly, wet state 2 (the 'wetter' state) also has a higher shape parameter in summer, combined with a higher scale parameter, suggesting this states is also capturing summer-time extreme rainfall. In wet state 1 (the 'wet' state), the probability of zero rainfall is higher in summer, leading to drier conditions in this state in this season. The rainfall variance and tail behaviour have more of a spring-autumn dipole, with the rainfall variance being higher in spring and lower in autumn, and vice-versa for the rainfall tail parameter. This suggests that, as well as it being less likely to rain in the spring compared to the autumn in this state, when



v

Figure 7. Graphical representations of the cyclic cubic regression spline fit to site V38, capturing the seasonal variation in each time-varying part of the HMM: (a) logit of the dry state persistence probability (Equation 3), (b) the logit of the zero rainfall probability in each state (Equation 4), (c) the log of the rainfall Generalised Pareto Distribution scale parameter in each state (Equation 5), and (d) the rainfall Generalised Pareto Distribution scale parameter in each state (Equation 6). The solid lines show the posterior median of the cubic regression splines, and the shaded regions show the 95% Bayesian credible intervals.

it does rain, the rain is more moderate, whereas in autumn it is more extreme. This state could therefore be thought to be representing the synoptic storm rainfall behaviour more associated with autumn and winter rainfall.

The 5-yearly random effects presented in Figure 8 show a reasonable amount of variability in the dry state persistence and rainfall scale parameter over the 105-year period. For example, the dry state persistence is shown to increase over the 1940s and then decrease again over the 1950s and 60s, peaking again in the early 1990s, while the GPD scale parameter experiences a considerable dip in the 1970s and 80s in wet state 1. These random effects are essentially 'mopping up' any variability in these

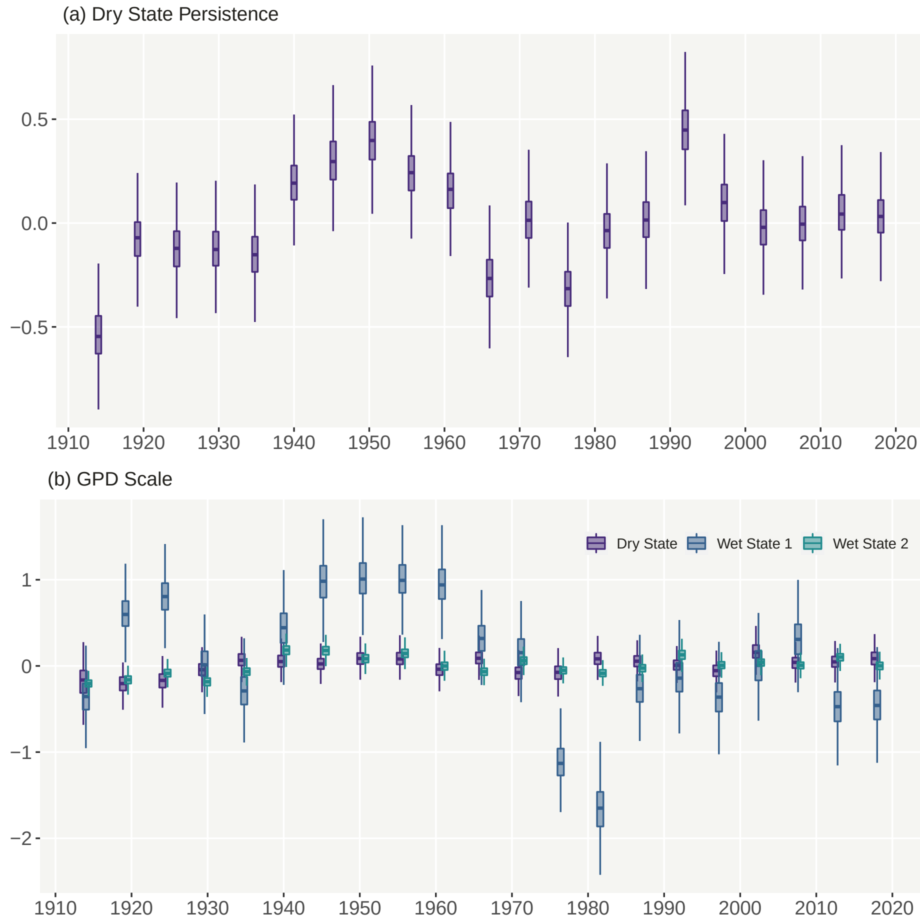


Figure 8. Graphical representations of the ‘time overall’ 5-yearly random effects fit to site V38, capturing the non-smooth temporal variability in (a) the logit of the dry state persistence probability (equation 3), and (b) of the log of the rainfall Generalised Pareto Distribution scale parameter in each state (equation 5). Each 5-yearly random effect is represented by a box and whisker plot, showing the 0, 0.25, 0.5, 0.75 and 1 quantiles of the posterior random effect parameter distributions as the lower whisker, lower box edge, middle box edge, upper box edge and upper whisker respectively.

rainfall characteristics, not captured by the climate drivers.

Figure 9 shows which of the possible 24 regression coefficients, representing the relationship between climate drivers and HMM model parameters π (the probability of no rain) and σ (the rainfall distribution scale parameter) in each state, are retained and which are ‘shrunk’ in each of the 39 site models in this application. This table shows how at a majority of sites, the EA pattern and SNAO have a non-zero effect on both π and σ in all three states, whereas the inclusion/exclusion of WNAO and SSTs varies more across sites. Figure 9 also shows the direction of each climate driver effect (i.e. positive or negative). These show how the models represent known relationships between climate drivers and rainfall behaviours. Namely how a positive EA index value is associated with wetter conditions, and hence a lower probability of zero rainfall and higher rainfall variability, while a positive SNAO index is associated with drier conditions, and hence a higher probability of zero rainfall and lower

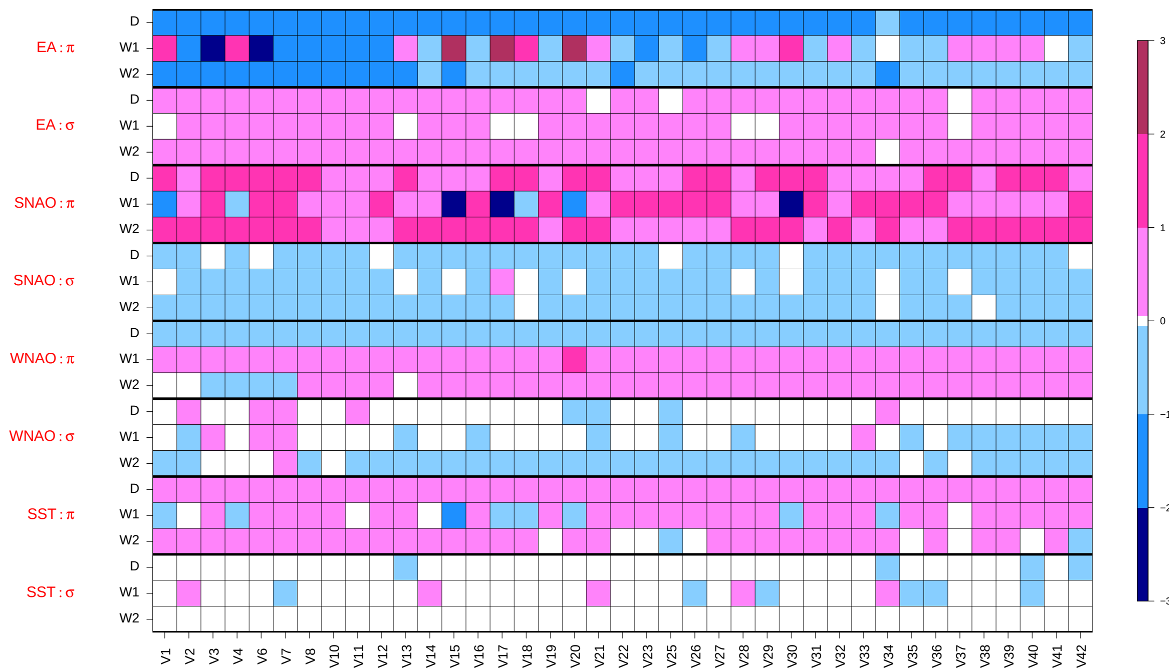


Figure 9. Table showing the posterior distribution mean of each of the 24 climate driver regression coefficients (y axis) in each of the rainfall HMMs fitted to the 39 sites (x axis). The colours indicate the direction and strength of the effect in each case, and those effects that are left blank/white are those that have been 'shrunk' away to zero by the Bayesian penalised regression when applied to that site (see Section 2.3). The red labels on the left side identify the relationship that each group of three rows is representing (e.g. EA:π represents the relationship between the EA pattern and π, the probability of zero rainfall), and each of the three rows is labelled as either D=Dry, W1=Wet or W2=Wetter, representing each state. For example, the box labelled (V1, W2, SST:σ) shows the posterior distribution mean of the regression coefficient explaining the relationship between SSTs and the rainfall distribution scale parameter (σ) in the 'wetter' state (W2) at site V1.

rainfall variability.

5.2.2 Validation of spatially coherent rainfall simulations

Figure 10 shows how, for all four sites and across all seasons, the distribution of observed daily rainfall is well represented by the associated simulated rainfall from the AME framework. The observed and simulated mean quantiles of the daily rainfall distributions lie close to the line of $y = x$, particularly for lower quantiles. In most cases, this line of $y = x$ falls within the Bayesian 95% credible interval of the simulated quantiles. Subsequently, when the observed and simulated daily rainfall are accumulated over each month within the 105-year period, a similar good consistency is seen between the simulations and observations at each of the four sites, as show in Figure 11. In order to also assess the performance of the model in the lower quantiles, plots equivalent to Figures 10 and 11, with log scale axes, are included in the Supplementary Material (Supplementary Figures 4 and 5). These plots show how the driest conditions are generally captured well, particularly on the daily scale, but that simulated monthly accumulations have a tendency to be too wet in the lower quantiles at site V16.

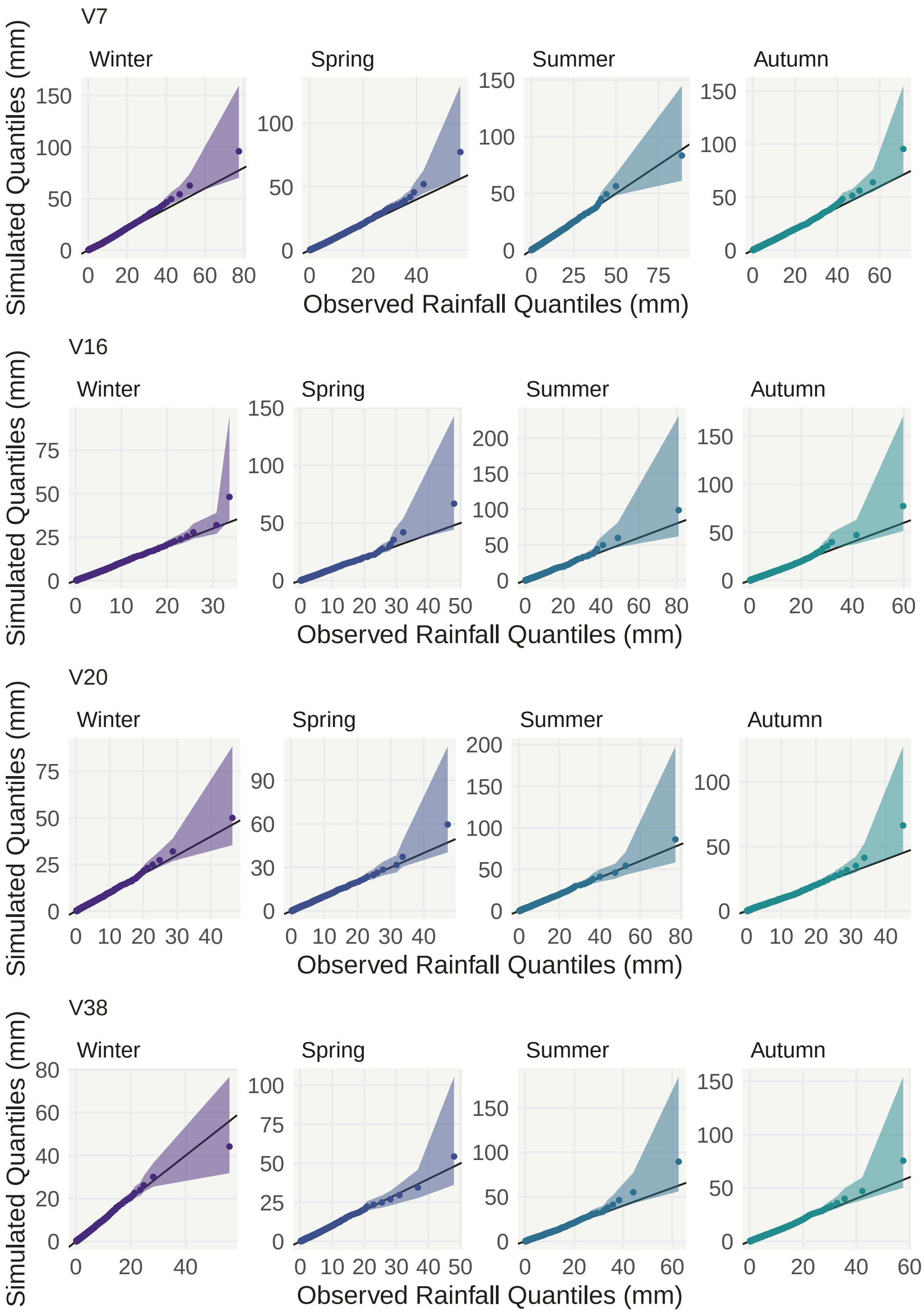


Figure 10. Quantile-quantile plots of the observed daily rainfall and AME simulated daily rainfall at four sites within the study region, spanning a range of rainfall climatologies, separated by season. The points in each plot show the observed quantile against the median of the quantile from the 1000 simulations from the AME framework, the shaded regions show the 95% Bayesian credible intervals for the quantiles from the 1000 simulations, and the solid black line is the line of y=x.

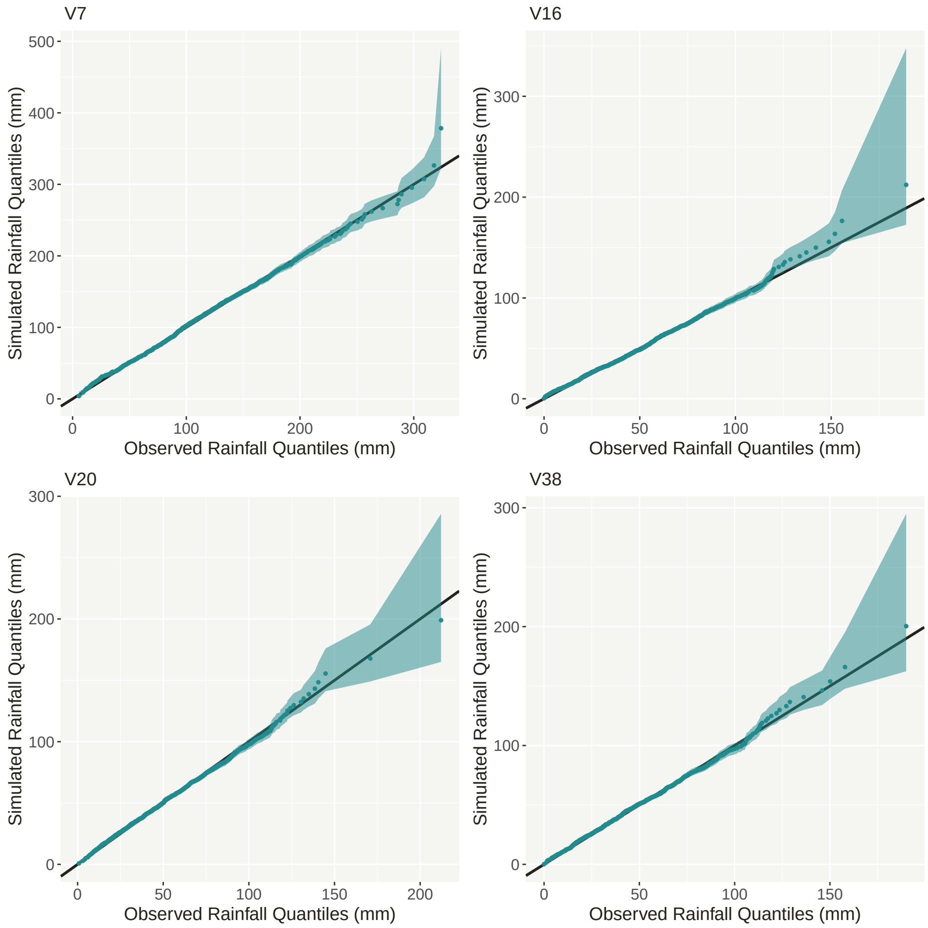


Figure 11. Quantile-quantile plots of the observed accumulated monthly rainfall and AME simulated accumulated monthly rainfall at four sites within the study region, spanning a range of rainfall climatologies. The green points show the observed quantile against the median of the quantile from the 1000 simulations from the AME framework, the green shaded region shows the 95% Bayesian credible interval for the quantile from the 1000 simulations, and the solid black line is the line of y=x.

In many of the plots in Figure 10, the upper bound of the 95% credible interval extends beyond the most extreme rainfall in the observed record. This is due to the heavy tailed shape of the observed rainfall distribution, such that most values are low with only a few much higher values. This means that the rainfall is best represented by a GPD distribution with a positive shape parameter, leading to an infinite upper tail to the fitted distribution and hence the potential for the simulation of high rainfall values. However, the daily rainfall simulations are all less than the realistic upper limit identified for the region (i.e. less than 300 mm), suggesting these simulated high rainfall days could plausibly occur.

Figure 12 shows good agreement between AME simulated and observed annual total rainfall quantiles (accumulated October-September) at these four sites. This is consistent across all 39 sites, as shown in Supplementary Figure 8. These plots are

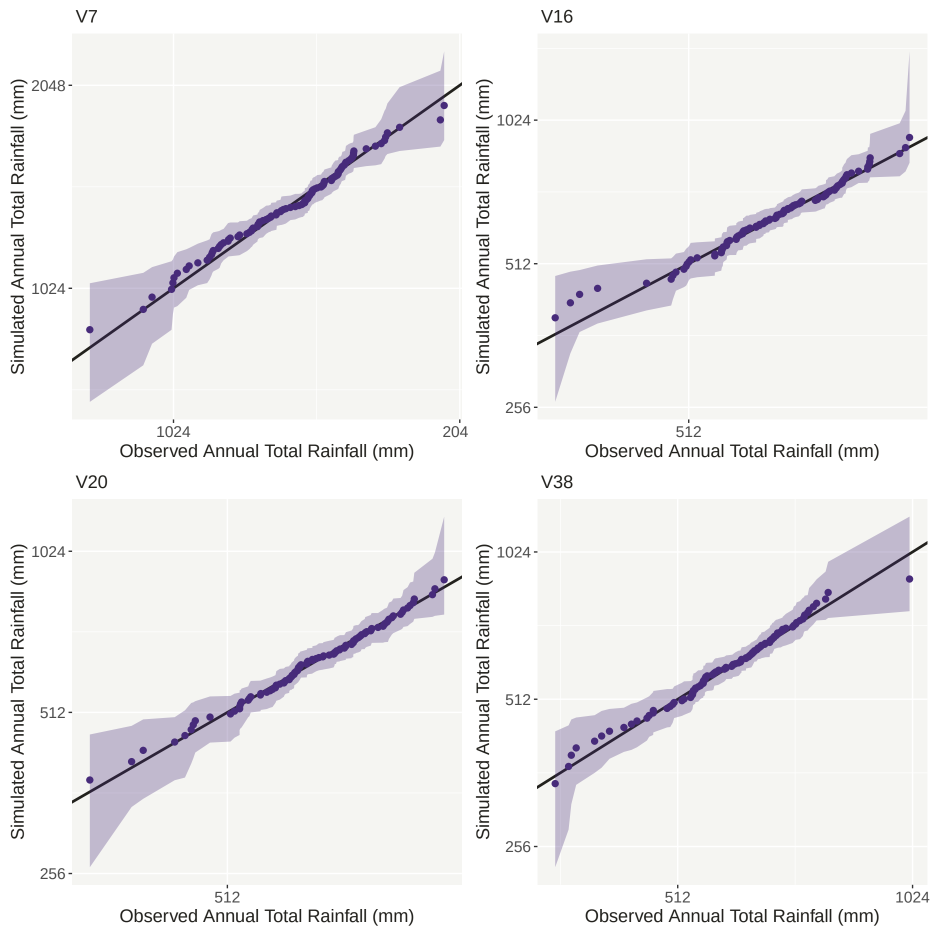


Figure 12. Quantile-quantile plots (on log scale axes) of the observed and AME simulated accumulated annual (October-September) rainfall at four sites within the study region, spanning a range of rainfall climatologies. The purple points show the observed quantile against the median of the quantile from the 1000 simulations from the AME framework, the purple shaded region shows the 95% Bayesian credible interval for the quantile from the 1000 simulations, and the solid black line is the line of y=x.

presented on a log scale to accentuate the lower quantiles (the driest years). Previous models developed and used within the UK water industry have been shown to systematically sample slightly wetter conditions on average in this lower tail, as noted in Section 1. These figures indicate that, while the AME framework shows this tendency in some locations (e.g. V16), for many others there is very good agreement (median points lying close to the line of y=x). This indicates that the AME framework is a step in the right direction in overcoming this, based on sound statistical theory.

The same good agreement between the observations and simulations can be seen in the proportion of zero rainfall days over the 105 years, as presented in Figure 13. In all sites and seasons, the median (0.5 quantile) of the 1000 simulations is very close to the observed proportion. This shows how the AME framework is able to faithfully capture the frequency of non-rainy days in the observed record, and how this varies with season, at each site.

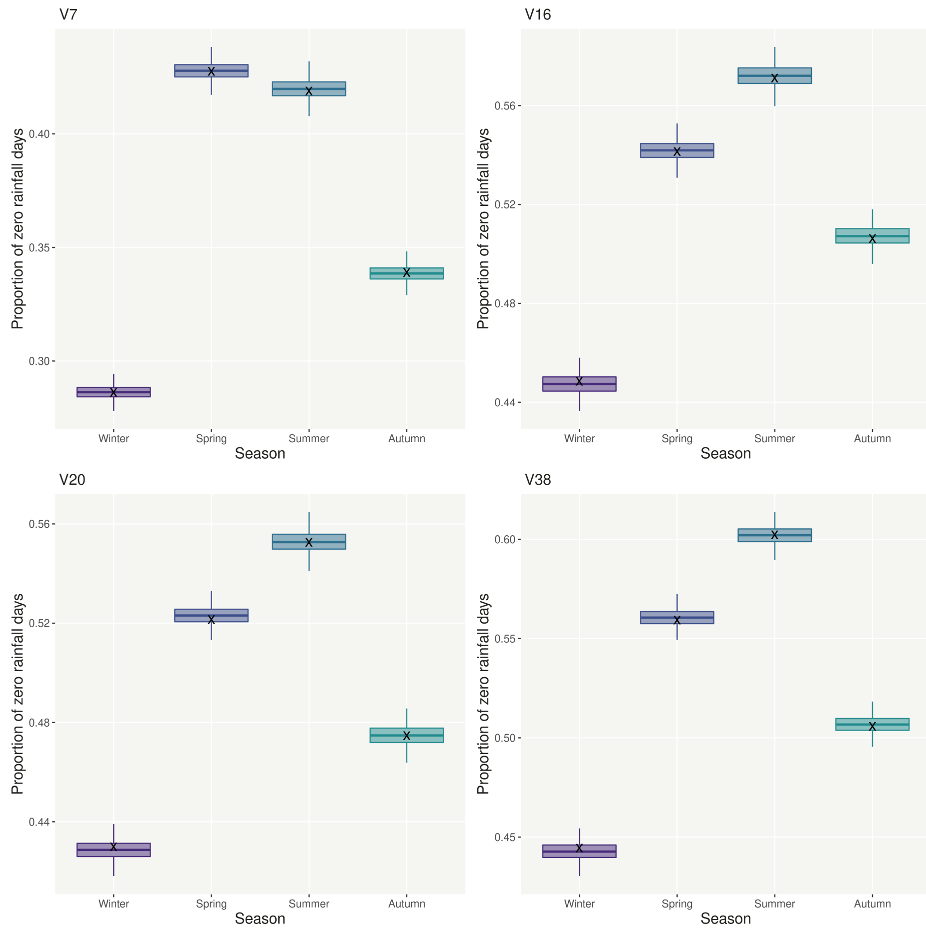


Figure 13. A comparison of the observed proportion of zero rainfall days in each season within the 105-year study period (black cross) and the same proportion in the 1000 AME simulations of the period (box and whisper plot), for four sites within the study region, spanning a range of rainfall climatologies. In each case the box and whisker shows the 0, 0.25, 0.5, 0.75 and 1 quantiles of the 1000 AME simulations as the lower whisker, lower box edge, middle box edge, upper box edge and upper whisker respectively.

Figure 14 shows how, for a majority of the range of observed dry period lengths, the distribution of the AME simulated dry period lengths matches closely to the equivalent distribution within the observed record. This is particularly true from sites V7, V16 and V20, and for the shorter dry period lengths. The AME simulated rainfall has a tendency to underestimate longer dry spells, particularly those greater than 20 days long. However, the line $y = x$ is just within the 95% credible interval of the AME simulations for the longest dry period observed at V38, showing how the AME is able to simulate such events, even if rarely. This component of the validation specifically tests short-duration (weekly to monthly) characteristics of the simulated rainfall.

Considering long-duration characteristics of rainfall, the observed and simulated rainfall behaviour is compared in terms of the 36-month accumulated rainfall DDI in Figures 15 and 16. Figure 15 shows how, at all four sites, the observed DDI is within the

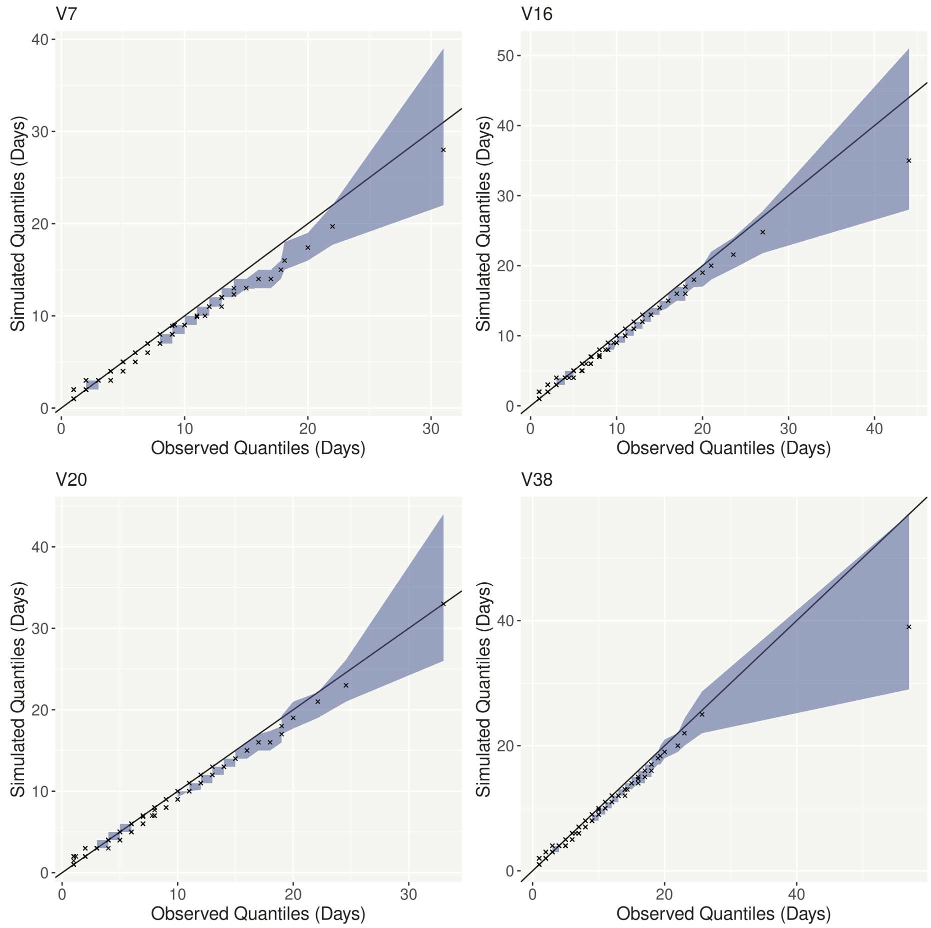


Figure 14. Quantile-quantile plots of the length of periods of zero rainfall days within the observations, and AME simulated rainfall series at four sites within the study region, spanning a range of rainfall climatologies. The black crosses show the observed quantile against the median of the quantile from the 1000 simulations from the AME framework, the blue shaded region shows the 95% Bayesian credible interval for the quantile from the 1000 simulations, and the solid black line is the line of y=x.

simulated range throughout the period. Indeed, the simulations are able to capture, and exceed, known very extreme drought periods such as those in the 1920s and 1930s at site V38. This is further shown in Figure 16, where the most extreme observed DDI values (associated with the greatest 36-month drought periods) are well captured by the median of the AME simulations. This figure and the equivalent plots for all 39 sites (Supplementary Figure 12) indicate that there is no systematic under- or over-estimation of long-duration drought severity.

While developing the model, it was found that extreme drought events were noticeably better captured when SSTs were added into the framework, highlighting the importance of their inclusion when aiming to represent long duration drought in the Greater Anglian region. As well as the observed DDI being within the simulated range, at all sites, the mean DDI from the simulations closely follows the variation seen in the index in the observed record. Again, this good agreement is thought to be due to the inclusion of relevant climate drivers, as well as the use of the most likely state sequence during simulation (see details of the simulation process presented in the Supplementary Material).

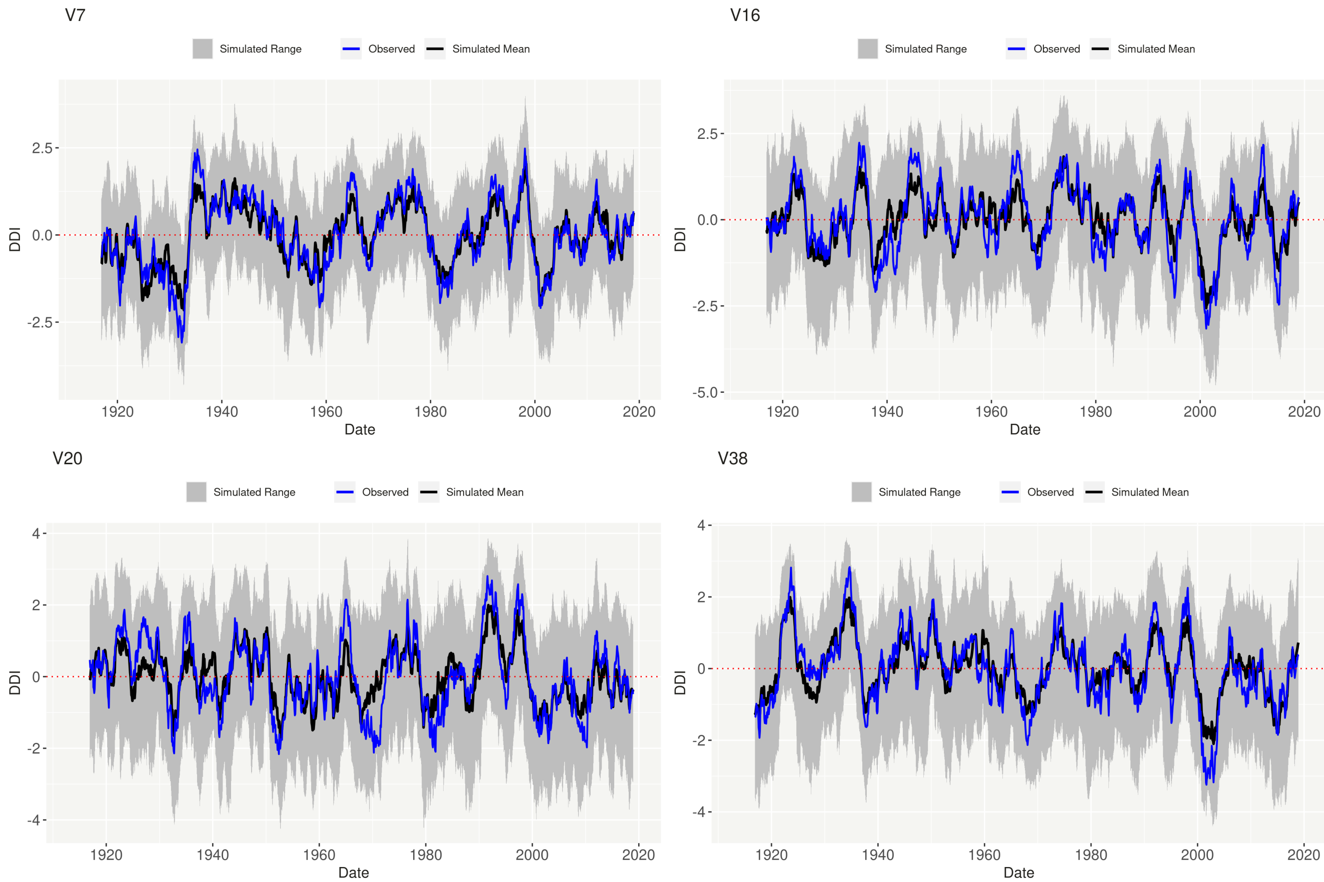


Figure 15. A comparison of the DDI time series based on 36-month running monthly rainfall accumulations from observed and from AME simulations for four sites within the study region, spanning a range of rainfall climatologies. The observed DDI is shown in blue, the range of the 1000 simulations is shown in grey and the mean of the simulations in black.

Note that Figures 10 - 16 validate the marginal distributions at each site, and not the spatial coherence. This spatial coherence is explored in Figures 17 and 18. Figure 17 shows how the observed between-site spatial dependence in monthly rainfall is generally well captured by the simulations, with low/high monthly rainfall at site V38 generally associated with low/high rainfall at other sites, as seen in the observations. The strength of the observed relationship between sites is less well captured by the simulations for sites that are closer together. For example, the bottom row of Figure 17 compares this relationship in the observed and simulated monthly rainfall at site V38 and the neighbouring site V41 (Woodbridge), showing a stronger correlation in the observations. This is further demonstrated in a systematic way in Figure 18, where the cross-correlations in daily and monthly rainfall for sites within approximately 150 km of each other are shown to be consistently slightly underestimated by

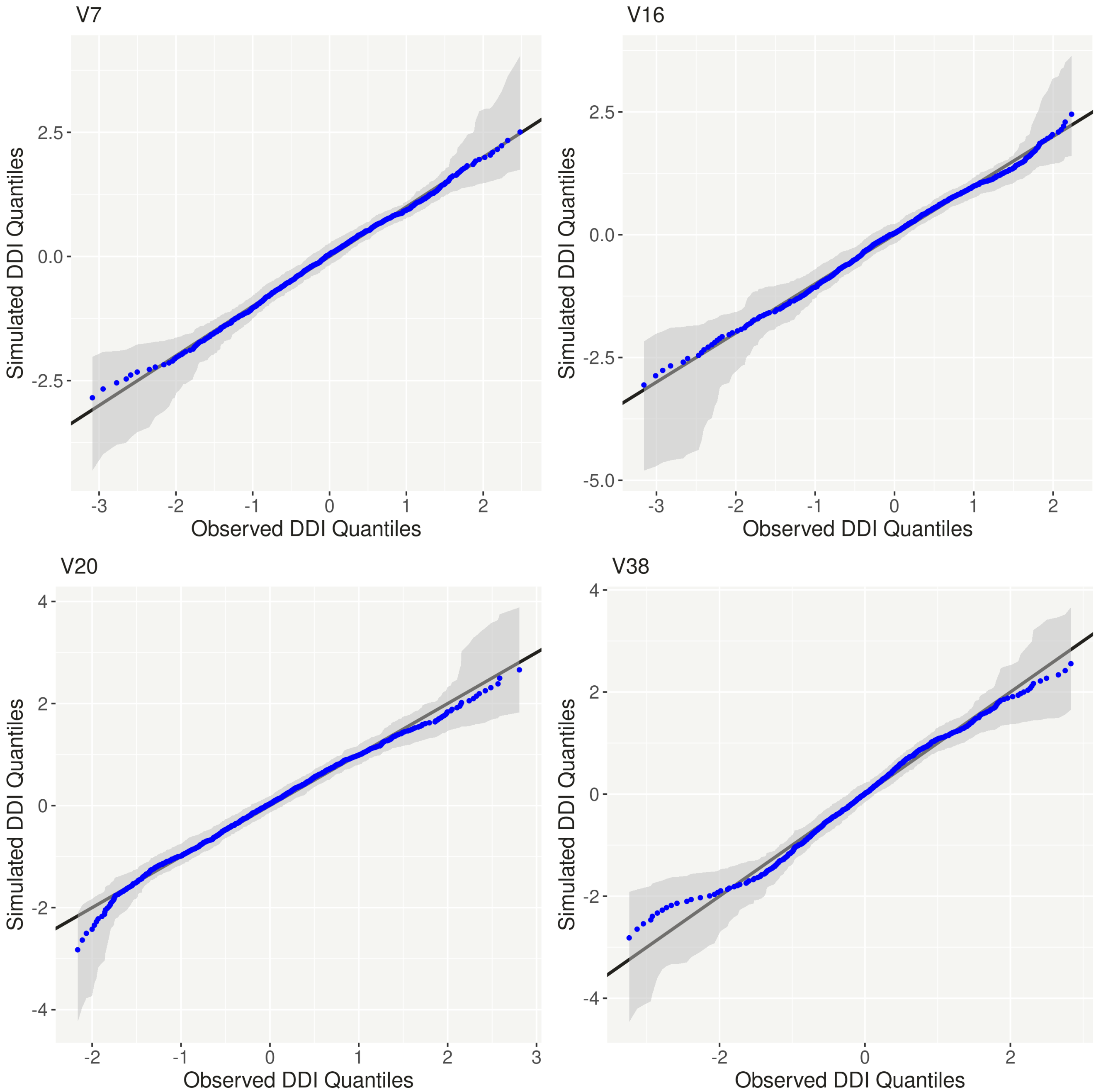


Figure 16. Quantile-quantile plots of observed and simulated DDI (based on 36-month running monthly rainfall accumulations) for four sites within the study region, spanning a range of rainfall climatologies. The blue points show the observed quantile against the median of the quantile from the 1000 simulations from the AME framework, the grey shaded region shows the 95% Bayesian credible interval for the quantile from the 1000 simulations, and the solid black line is the line of y=x. Note that the upper tail of the DDI distribution relates to dry conditions.

the AME simulations.

It may be hypothesised that this misrepresentation of the dependence between locations could be due to the regular vine (rather

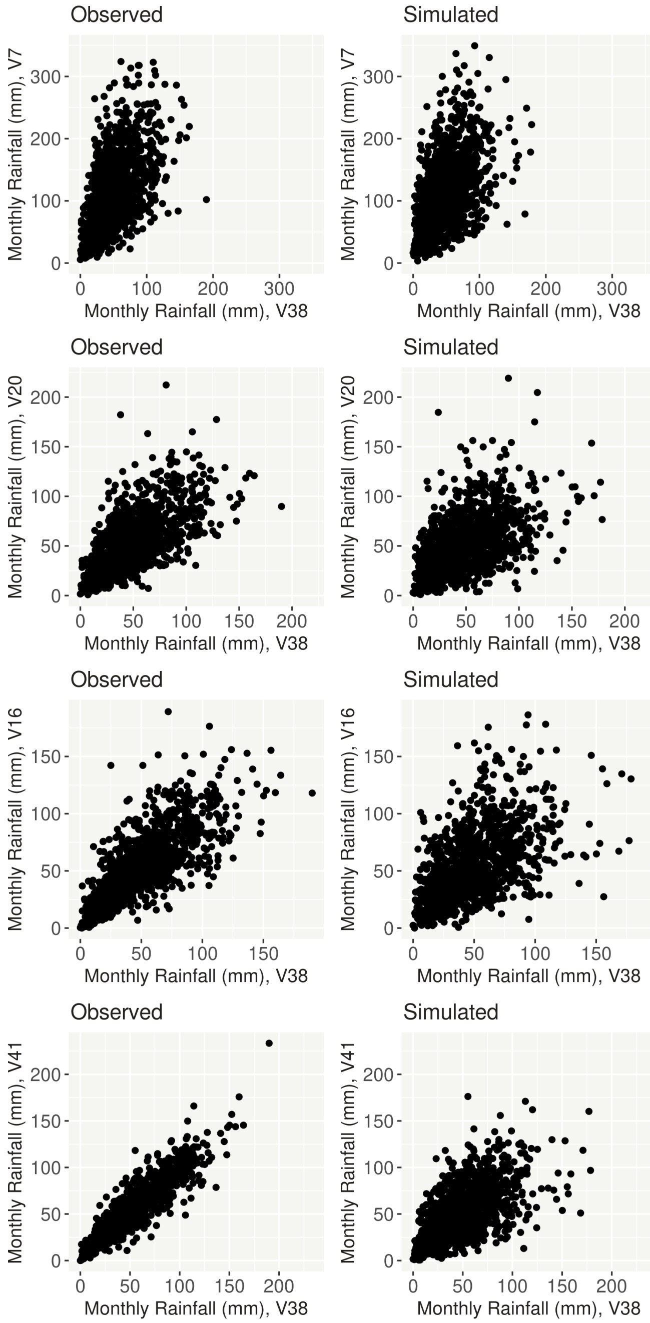


Figure 17. A comparison of the observed (left column) and 1 AME simulated (right column) pairwise relationship between monthly accumulated rainfall at site V38 and four other sites separated by varying distances in space (V38-V7:243km, V38-V20:182km, V38-V16:122km, V38-V41:30km).

than fully cross-correlated) dependence structure of the vine-copula used within the AME framework. This is, however negated by observing that the pair of sites shown to be most misrepresented in Figure 17 (V38 and V41), are a pair that are explicitly modelled using a bivariate Gumbel copula in Tree 1 of the vine-copula (nodes 35 and 38 in Figure 5). Rather, this small inconsistency in the simulations is concluded to be a result of the methodology necessitating the independent modelling of

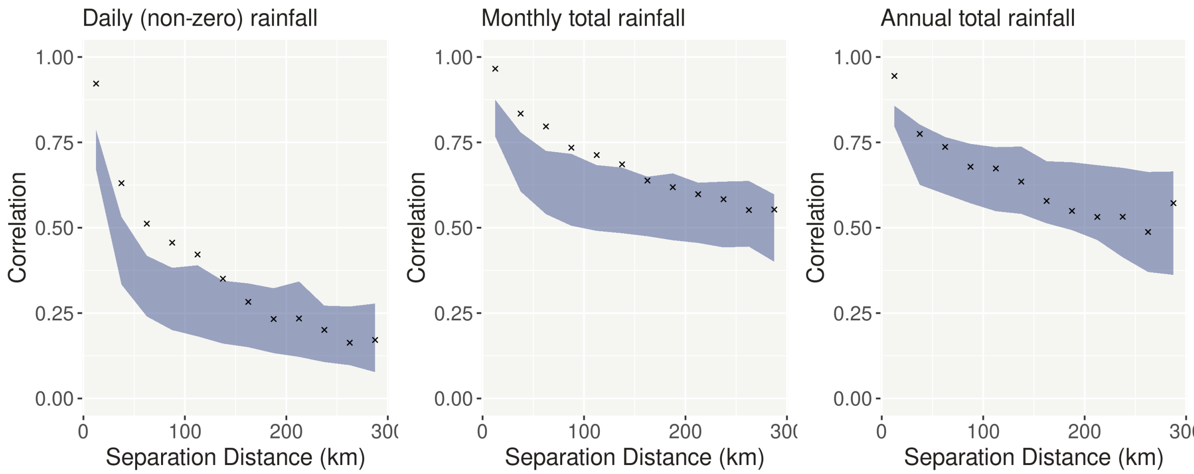


Figure 18. A systematic comparison of the observed and AME simulated pairwise cross-correlations in (non-zero) daily, monthly total and annual total rainfall, plotted as a function of separation distance. Within the observations and each of the 1000 AME realisations, the mean cross correlation in each 25km separation distance 'bin' is calculated. The black crosses show this mean value against the mid-point of the separation bin for the observed data. The shaped region shows the range of equivalent values calculated from the 1000 AME realisations.

hidden state sequences at each site, as discussed in Section 2.4. Hence, even if the spatial dependence in the residual rainfall is represented perfectly by a copula, using non-spatially coherent hidden states across the sites will result in the residual rainfall being transformed to the rainfall scale incoherently. For example, suppose a rainfall residual value of 0.8 is simulated at a location, this could be transformed to say 10mm in the wet state or 25mm in the very wet state because of the differences in the rainfall distributions in the two states. If the model is not explicitly capturing how nearby sites' states co-vary, this could result in a mismatch in the subsequent simulated rainfall, as shown in Figures 17 and 18. As described in Section 2.4, this methodological compromise is necessary to achieve a computationally feasible modelling framework, and these figures demonstrate that this compromise has only a small impact on the model's performance. Further, this is found to have very little impact on the final gridded simulated output (see Section 5.2.3).

5.2.3 Validation of gridded rainfall

An important next step in the validation process is to extend on the work of Section 5.2.2 and validate the gridded datasets produced at step 6 of the methodological diagram shown in Figure 1. Three key tests are covered (metrics 11-13 in Section 5.1), the results of which are presented here.

Greater Anglian region validation

Calculating the monthly mean rainfall (across all months in the 1914-2018 period) and comparing the observed and simulated spatial variability in the Greater Anglian region serves as an initial validation. Note that this is a simple check of AME performance. This does not, for example, test the spatial dependence structure. Any simple model, or interpolation technique, should be able to reproduce the spatial variability of annual mean rainfall and that is the purpose of this check. The model is able to simulate the observed spatial gradients, clearly depicting the driest and wettest sub-regions within the simulated domain (Figure 19). Reassuringly, errors in the climatology across the region are small (typically less than 5%) and show no clear tendency towards positive or negative values in a particular sub-region (Supplementary Figure 13). The fact that errors are small is expected given that the extension to a grid (Section 2.5) uses gridded long-term average rainfall in its approach. Note that although two AME realisations are chosen at random, these results are robust to any selection of AME realisations (not shown).

The Greater Anglian region average DDI is calculated using 12-, 24- and 36-month rainfall accumulations. The 36-month DDI is shown in Figure 20, which indicates that the observed variability in this drought index is well captured by the AME model. Where the AME model mean DDI (the thick black line) deviates slightly from the observed HadUK-Grid DDI (thick blue line), the observed behaviour is still captured by the range of the 1000 AME realisations (thin grey lines). The same is true for both 12- and 24-month DDI (not shown).

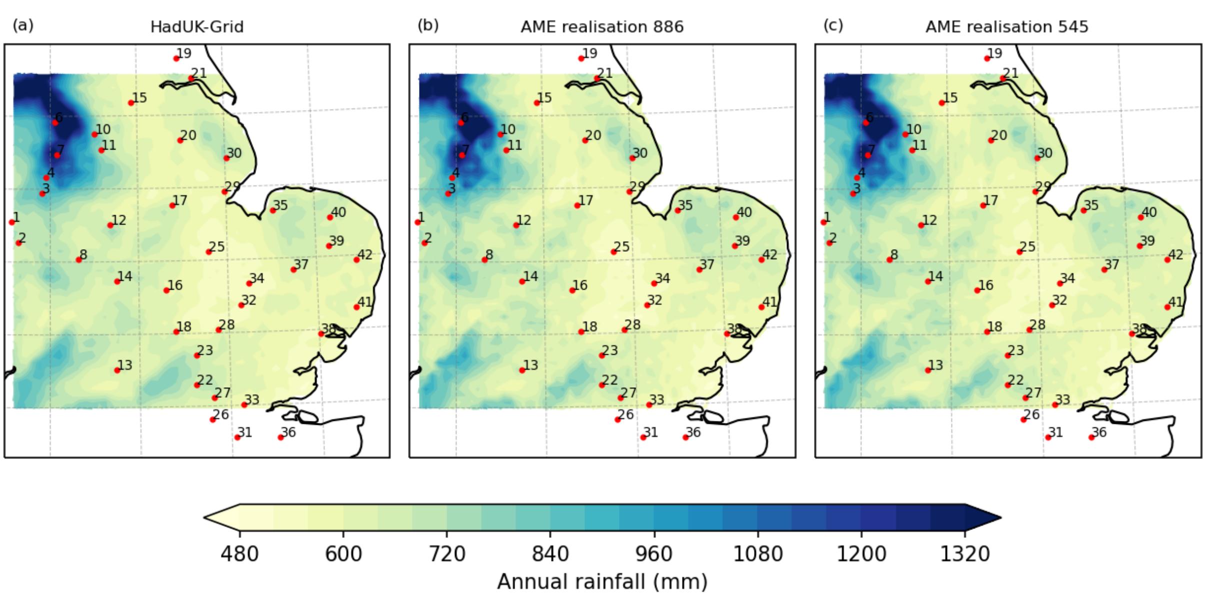


Figure 19. Annual mean rainfall (mm) (using a 1961-1990 climatology period) across the simulated domain in HadUK-Grid and two (of 1000) randomly chosen AME realisations. The 39 sites are labelled.

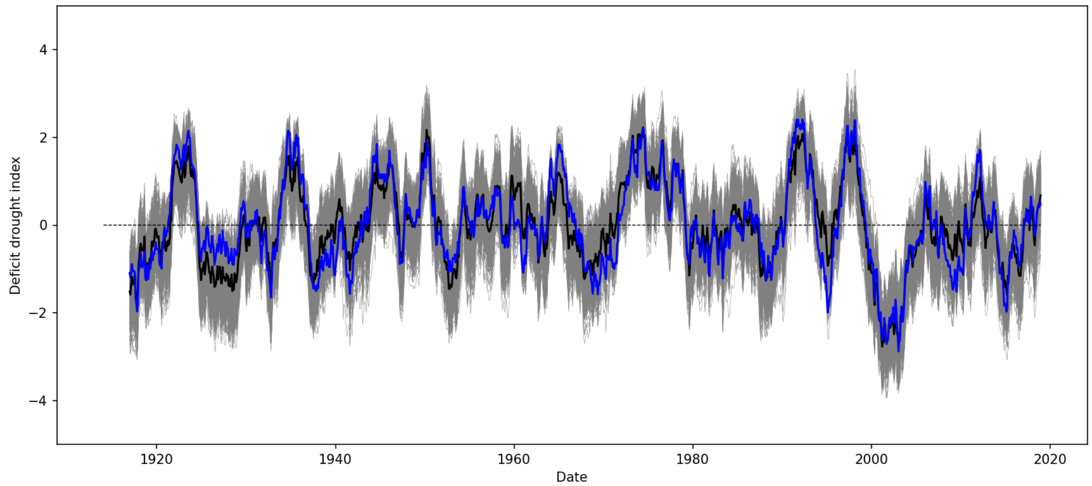


Figure 20. DDI using 36-month running monthly rainfall accumulations for the Greater Anglian region. The thin grey lines represent all 1000 AME simulations, the thick black line represents the average of the 1000 AME simulations and the thick blue line represents HadUK-Grid.

Independent site validation

At Writtle, the observed 12-, 24- and 36-month non-standardised DDIs are well captured by the AME model (Figure 21). Recall that this validation is more a test of the interpolation technique that produces the gridded dataset, rather than the direct simulation of data at the 39 sites (see metrics 12 in Section 5.1). There is a slight tendency to underestimate the magnitude of the most extreme meteorological droughts. This is also illustrated in Figure 22, which shows that there is a slight underestimation of the magnitude of 36-month meteorological drought events. However, all but the most extreme observed droughts (in magnitude, severity and duration) are within the range of AME model simulations. This represents good performance for an independent site that is some distance from the nearest modelled sites. Further, the agreement between simulated and observed 12- and 24-month meteorological droughts is greater (relative to 36-month droughts) (indicated by Figure 21). Very similar results are found for the Shoeburyness site.

For Cromer the results are slightly different. The AME mean DDI generally closely tracks the observed DDI, although some variability later in the period is not as well represented, such as the wetter period prior to 2010 (Supplementary Figure 15). This wet period seems to be restricted to the northeast of the Greater Anglian region and is not so apparent in observations

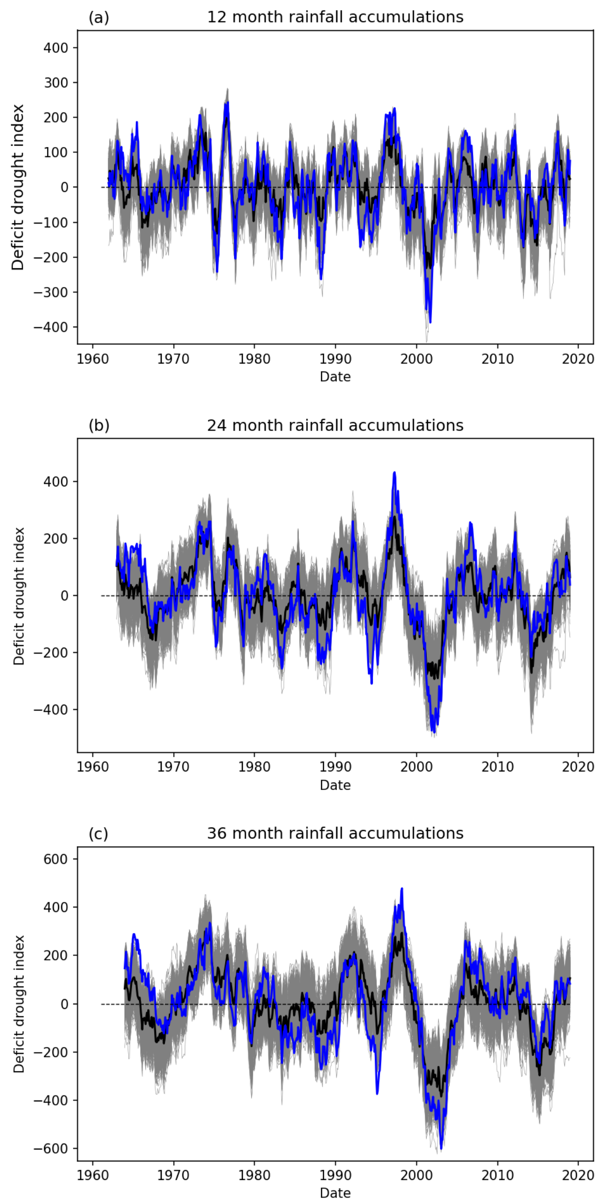


Figure 21. The non-standardised DDI for Writtle using (a) 12-, (b) 24-, and (c) 36-month running rainfall accumulations. The thin grey lines represent all 1000 AME simulations, the thick black line represents the mean of the AME simulations and the thick blue line represents the observations, using HadUK-Grid.

elsewhere. If it was a local feature it may not be captured by the 39 sites. Generally, there is greater variability across the 1000 AME simulations, relative to, for example, Writtle (cf. Figure 21 and Supplementary Figure 15). In turn, the range across the 1000 AME simulations encompasses the observed meteorological drought characteristics. Very similar results are apparent for Wattisham, Brooms Barn and Monks Wood (not shown).

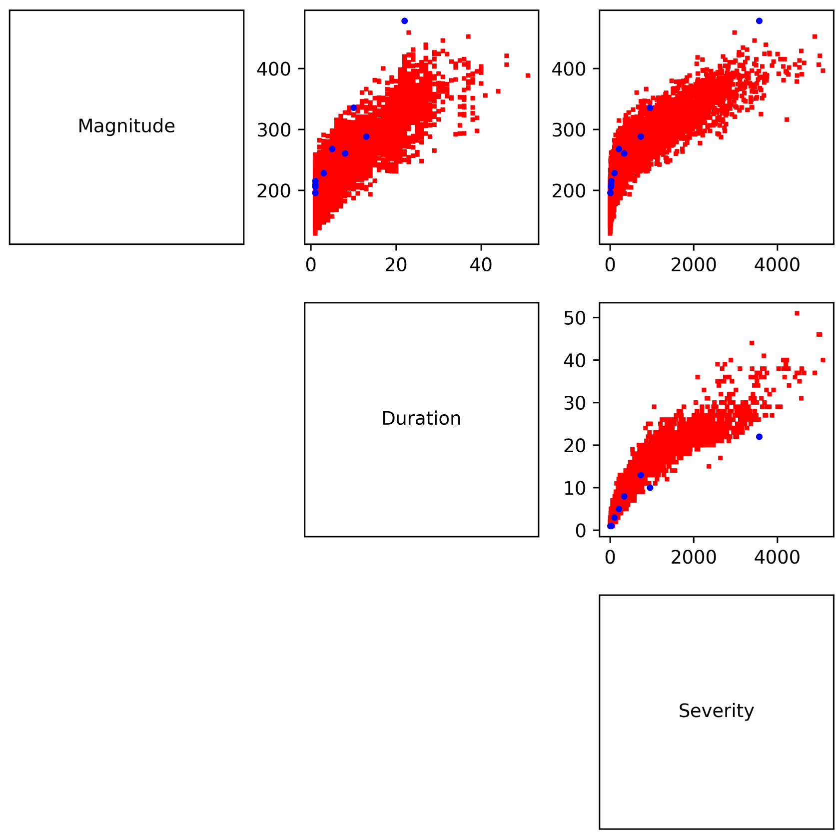


Figure 22. Comparing 36-month meteorological drought characteristics of magnitude (mm) duration (month) and severity (mm $\times$ month) for observed rainfall at Writtle (blue circles) and 1000 AME model simulations (red circles). Meteorological drought events are identified as periods that exceed the 90th percentile of the DDI, with an event terminating once it drops below this threshold.)

All sites demonstrate improved performance of simulated meteorological droughts at 12- and 24-month durations compared to 36-month durations (not shown), tending to slightly underestimate the absolute magnitude of rainfall accumulation deficits and surpluses (likely due to locally significant events that are not captured by rain gauges at the nearest of the 39 sites). Overall, this slightly dampened month-to-month rainfall variability away from the modelled sites is an issue with interpolated gridded datasets that are generated with a finite number of sites. Improving the representation of rainfall at a location further away from one of the 39 sites is difficult because the true rainfall is not always known (only limited additional site data beyond the long daily rainfall records at the 39 sites is available).

Validation in seasons and key periods

Considering maps of daily rainfall (aided through the production of animations) many features appear in both HadUK-Grid and the AME model. For example, runs of both dry and wet days are apparent in both, with occasional days of very wet weather across the region. The lack of seasonality in rainfall across the region prohibits strong comparisons between summer and winter. There does not appear to be a noticeable difference in the average pattern of rainfall in summer and winter in either HadUK-Grid or the AME model; it is not possible to clearly differentiate predominantly frontal or convective rainfall days (this is a limitation of the simulating at daily frequency, but a relatively minor one for the purpose of drought investigation).

The daily rainfall simulated by the AME model often appears "spotty". This is related to the fact that spatially coherent rainfall is simulated at the 39 sites and then interpolated. As a consequence, rainfall does not always appear as homogeneous as it might be in reality. Clearly, rainfall is a heterogeneous meteorological variable, especially at the daily time-scale. Therefore, it is difficult to pinpoint exactly how much of this "spotty" behaviour is due to this inherent feature of rainfall and how much is due to a limited number of sites that the AME model is fitted to. Indeed, this "spotty" behaviour is still apparent in HadUK-Grid, even in winter when frontal rainfall is more common, suggesting it does not purely represent convective rainfall but rather a potential limitation of all gridded rainfall datasets. However, this feature is arguably less pronounced in HadUK-Grid relative to the AME model, which is expected given that more than 39 sites will underpin the daily rainfall across the region on most days in HadUK-Grid; the exact number of sites available will vary in time and, unlike in the AME model, will not be fixed at a particular number of sites.

6 Discussion and Conclusion

A novel stochastic rainfall model, named the Advanced Meteorology Explorer (AME), has been presented, and its ability to capture observed rainfall and meteorological drought behaviour in the Greater Anglian region of the UK has been validated. In particular, the AME framework was developed to overcome the limitations of the current UK water industry standard stochastic rainfall model, as described in Section 1. This framework provides daily rainfall simulations on a high-resolution grid, capturing the effect of important climate drivers in a flexible way, within a Bayesian hierarchical framework.

The AME represents daily rainfall behaviour at a number of individual locations using a series of advanced HMMs, allowing for complex rainfall patterns to be capture through the hidden state structures which include clone dry states. Improved model performance is achieved by conditioning the HMM parameters on both temporal and climate driver effects. This is done in a flexible way, whereby different climate drivers are able influence rainfall at different sites, selected based on Bayesian penalised regression. The dependence between sites at different locations in space is modelled using a vine-copula, able to capture both asymptotic extremal dependence and independence between locations. The resulting spatially and temporally coherent rainfall simulations at the modelled sites are then interpolated to a 5 km grid using the same terrain dependent inverse-distance weighting approach used to create the HadUK-Grid data set (Hollis et al., 2019).

The AME framework is fit to daily rainfall data from the HadUK-Grid dataset in 39 1 km grid cells, representative of rain gauge sites in the Greater Anglian region of the UK, during the period 01/01/1914-31/12/2018. The resulting model is used to simulate 1000 alternative synthetic realisations of this 105-year period at the 39 sites, which are subsequently interpolated to a 5 km grid over the region of interest.

The validation of these stochastic simulations showed how, at the modelled sites, the AME framework is able to capture well the distribution of dry period lengths; seasonal and annual variation in occurrence and intensity of rainfall; and the distribution of rainfall intensity when aggregated to monthly and yearly resolutions. This is shown to be true for both very dry (V38) and relatively wet (V7) sites within the region. In addition, the simulated rainfall is shown to be within a realistic range. Subsequently, the AME framework is shown to represent long-duration (36-month) meteorological drought behaviour at modelled sites well. The mean DDI, calculated over the 1000 stochastic simulations, is shown to follow the variations in the observed index closely, and a number of stochastic simulations provide more extreme values than those seen in the most extreme meteorological droughts at the driest sites (V38).

There are three key limitations of the AME framework. Firstly, the high computational cost associated with applying the complex HMM within a Bayesian modelling framework. Secondly, the desired close match to the observed rainfall behaviour and subsequent high dependence of the simulations on the observations (which are just one single realization of the process), which might ultimately lead to the underestimation of the true variability of the stochastic process. Thirdly, the underestimation of spatial correlation in rainfall, particularly for locations in close proximity. To allow for computational feasibility when applying the AME framework to a large number of sites (here 39), the spatial dependence between sites is modelled separately from the magnitude and occurrence of rainfall, following a copula approach. The copula is fit to the residual rainfall, left after transforming the rainfall at each site to the uniform scale using the HMM distribution function at that site, and integrating out the hidden state sequence. Since the state sequence must be integrated out, this part of the modelling framework is represented independently at each site, and hence the dependence between hidden states at different sites is not explicitly captured. As a result, the correlation between daily, and monthly rainfall accumulations at nearby sites is shown to be slightly less strong in the simulations from the AME framework compared to the observations. As described in Section 2.4, methods based on coupled HMM have been proposed, but as currently formulated they are not computationally feasible for our use case, where the number of sites would result in the estimation of a huge number of additional state transition parameters, limiting the modelling to just a few sites. The approach taken in this study could, instead, be applied to even more sites. Further exploration would be required to determine an upper limit to the number of sites, but the computationally efficient fitting of the vine-copula model (taking just 8 minutes in this 39-dimensional example) suggests that more than 100 sites could be comfortably modelled together in this AME framework. Further, the way in which the AME models state sequences at each site conditional on the same temporal covariates, and based on spatially coherent rainfall data, means that some dependence in the state sequences is indirectly inferred. As such, this necessary compromise in the framework is shown to have little impact on the final gridded simulated output.

Indeed, the gridded output is shown to validate convincingly at both the regional scale (taking regional averages) and at several independent rainfall sites, representing sub-regions not covered by the 39 modelled sites (e.g. Essex). While the absolute magnitude of meteorological droughts at these independent sites is shown to be slightly dampened, tending towards climatology (common feature of gridded rainfall datasets), key characteristics of observed droughts are simulated. In addition, the AME is shown to produce plausible, more severe droughts than seen in the observed record.

The AME approach presents multiple improvements over the current industry approach (described in Section 1), drawing upon the advantages of many similar models in the literature and extending them to enhance model flexibility. The use of a spatial interpolation method to extend to a high-resolution 5 km dataset also represents a significant advancement on previous industry work. Although there is some relaxation towards climatology and an underestimation of monthly rainfall variance at independent sites, this is a common feature of gridded rainfall datasets. The interpolation scheme used here is designed specifically for daily rainfall and allows for direct comparisons with the observed HadUK-Grid dataset, since it uses an identical approach (Hollis et al., 2019). These features represent an improved ability to produce faithful hydrological modelling in the Greater Anglian region, particularly so at the catchment scale.

In this application, the simulations from the AME framework are constrained by the observed behaviour of the climate drivers during the historical 105-year simulation period. That is, the same values of SNAO, WNAO, EA and SST are used in the simulations as are observed in the model fitting period. This means that the simulated rainfall time series may not sample the full range of plausible climate scenarios. To increase this range, a synthetic or climate model simulated sequence of these climate drivers could also be used for simulation. This would allow the AME framework to sample a larger range of plausible climate variability, and hence plausible droughts.

In addition, the AME framework has the potential to be extended in a number of ways, similar to comparable models in the literature. For example, the same method could be used to model rainfall in another region of interest. The flexibility of the model in capturing climate driver effects, and its success in modelling both dry (V38) and wet (V7) sites in the Greater Anglian region, suggests that the AME framework could successfully represent locations and regions with very different rainfall climatologies to those seen in this application. Interpolation to the grid, may however be less successful in very hilly regions, where the terrain adjustment in the interpolation method may not be able to capture steep changes in elevation.

Further, while in this application the focus of the validation has been on characterising meteorological drought, the simulations from the AME framework could be used to explore infrastructure flooding in the insurance industry, waste water management, and any industry or application that is impacted by rainfall. While further model validation would be required to explicitly verify the AME's ability to realistically capture the rainfall characteristics relevant for each application, the overall good performance of the model in representing the full distribution of rainfall at different sites (including high rainfall values) shows good promise in its transferability to other applications.

The framework also has the potential to be extended to represent and simulate additional meteorological variables, such as Potential Evapotranspiration (PET) or temperature. This could be done in a similar way to modelling multiple locations, using a copula to represent the dependence between residual rainfall and (say) PET at each modelled location. Alternatively, a 2-dimensional HMM could be implemented at each location if found to be computationally feasible.

Finally, since the AME framework has been developed to assess environmental risk, to inform *future* infrastructure planning in the water industry, a quantification of the effect of future climate change on rainfall behaviour is also important. In future work we plan to extend the AME framework to incorporate information from the latest UK climate projections (UKCP), released by the Met Office in 2018 (Lowe et al., 2018). These projections indicate that, in general, summers will become drier while winters will become wetter, impacting on both drought and flood risk. This extension of the AME framework will aim to use an approach similar to Brown et al. (2014), and model both historical observations and climate model projections together, capturing how the rainfall behaviour (i.e. HMM parameters) changes with climate change at each site by regressing on global mean temperature. This will allow for stochastic simulations of rainfall, representative of any global mean warming level of interest.

In conclusion, the AME framework has been shown to be able to meet the needs of the UK water industry in simulating spatially and temporally coherent synthetic daily rainfall on a high-resolution grid, with no post-processing required to faithfully represent the observed characteristics of rainfall and resulting meteorological drought even when aggregating over spatial domains, as is the current industry approach. Indeed, the simulations are able to capture and exceed the extreme drought conditions experienced in historical record. The framework has the potential to be extended in many different ways, such as to include the effect of climate change, and rainfall simulations from the model can be used for a wide range of applications beyond water resource management.

References

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

Ansell, T. J., Jones, P. D., Allan, R. J., Lister, D., Parker, D. E., Brunet, M., Moberg, A., Jacobeit, J., Brohan, P., Rayner, N. A., Aguilar, E., Alexandersson, H., Barriendos, M., Brandsma, T., Cox, N. J., Della-Marta, P. M., Drebs, A., Founda, D., Gerstengarbe, F., Hickey, K., Jónsson, T., Luterbacher, J., Nordli, O., Oesterle, H., Petrakis, M., Philipp, A., Rodwell, M. J., Saladie, O., Sigro, J., Slonosky, V., Srnec, L., Swail, V., García-Suárez, A. M., Tuomenvirta, H., Wang, X., Wanner, H., Werner, P., Wheeler, D., and Xoplaki, E. (2006). Daily Mean Sea Level Pressure Reconstructions for the European–North Atlantic Region for the Period 1850–2003. *Journal of Climate*, 19(12):2717–2742.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52:345–370.

Breinl, K., Turkington, T., and Stowasser, M. (2013). Stochastic generation of multisite daily precipitation for applications in risk management. *Journal of hydrology*, 498:23–35.

Breinl, K., Turkington, T., and Stowasser, M. (2015). Simulating daily precipitation and temperature: a weather generation framework for assessing hydrometeorological hazards. *Meteorological Applications*, 22(3):334–347.

Brooks, N. (2012). The August 1912 floods in Norfolk. *Weather*, 67:204–205.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(2):434–455.

Brown, S. J., Murphy, J. M., Sexton, D. M. H., and Harris, G. R. (2014). Climate projections of future extreme events accounting for modelling uncertainties and historical simulation biases. *Climate Dynamics*, 43(9–10):2681–2705.

Buishand, T. A. and Brandsma, T. (2001). Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. *Water Resources Research*, 37:2761–2776.

Burke, E. J., Perry, R. H., and Brown, S. J. (2010). An extreme value analysis of UK drought and projections of change in the future. *Journal of Hydrology*, 388(1):131–143.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97:465–480.

Chandler, R. (2019). *RGLIMCLIM: A multisite, multivariate daily weather generator based on Generalized Linear Models, User Guide*. Department of Statistical Science, University College London, Gower Street, London, UK.

Coles, S. G., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analysis. *Extremes*, 2:339–365.

Czado, C. (2019). *Analyzing Dependent Data with Vine Copulas*. Springer Nature, Switzerland.

Dawkins, L. C. and Stephenson, D. B. (2018). Quantification of extremal dependence in spatial natural hazard footprints: independence of windstorm gust speeds and its impact on aggregate losses. *Natural Hazards Earth System Sciences*, 18:2933–2949.

Dissmann, J. F., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics and Data Analysis*, 59(1):52–69.

Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Fereday, D., O'Reilly, C., Stirling, A., Eade, R., Gordon, M., MacLachlan, C., Woollings, T., Sheen, K., and Belcher, S. (2018). Skilful seasonal predictions of summer European rainfall. *Geophysical Research Letters*, 45(7):3246–3254.

Folland, C. K., Knight, J., Linderholm, H. W., Fereday, D., Ineson, S., and Hurrell, J. W. (2009). The summer North Atlantic Oscillation: Past, present, and future. *Journal of Climate*, 22:1082–1103.

Francis, T. (2011). Extreme Precipitation Analysis at Sizewell: Final Report.

Furrer, E. M. and Katz, R. W. (2007). Generalized linear modeling approach to stochastic weather generators. *Climate Research*, 34:129–144.

Gelman, A., Carlin, J. B., H. S. Stern, D. B. D., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis (Third Edition)*. Chapman & Hall/CRC, Boca Raton, Florida, USA.

Hall, R. J. and Hanna, E. (2018). North Atlantic circulation indices: links with summer and winter uk temperature and precipitation and implications for seasonal forecasting. *International Journal of Climatology*, 38(S1):660–677.

Hollis, D., McCarthy, M., Kendon, M., Legg, T., and Simpson, I. (2019). HadUK-Grid: A new UK dataset of gridded climate observations. *Geoscience Data Journal*, 6:151–159.

Holsclaw, T., Greene, A. M., Robertson, A. W., and Smyth, P. (2016). A bayesian hidden markov model of daily precipitation over south and east Asia. *Journal of Hydrometeorology*, 17(1):3–25.

Huser, R. G. and Wadsworth, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *Journal of the American Statistical Association*, 114(525):434–444.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20(1):50–67.

Katz, R. W. (1977). Precipitation as a chain-dependent process. *Journal of Applied Meteorology*, 16:671–676.

Katz, R. W. and Parlange, M. B. (1998). Overdispersion phenomenon in stochastic modeling of precipitation. *Journal of Climate*, 11:591–601.

Kleiber, W., Katz, R. W., and Rajagopalan, B. (2012). Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water Resources Research*, 48(1):W01523.

Kornhuber, K., Osprey, S., Coumou, D., Petri, S., Petoukhov, V., Rahmstorf, S., and Gray, L. (2019). Extreme weather events in early summer 2018 connected by a recurrent hemispheric wave-7 pattern. *Environmental Research Letters*, 14(5).

Kroiz, G. C., Majumder, R., Gobbert, M. K., Neerchal, N. K., Markert, K., and Mehta, A. (2020). Daily precipitation generation using a hidden markov model with correlated emissions for the Potomac river basin. *Proceedings in Applied Mathematics and Mechanics*, 20(1):e202000117.

Legg, T. (2015). Uncertainties in gridded area-average monthly temperature, precipitation and sunshine for the United Kingdom. *International Journal of Climatology*, 35(7):1367–1378.

Li, C., Singh, V. P., and Mishra, A. K. (2012). Simulation of the entire range of daily precipitation using a hybrid probability distribution. *Water Resources Research*, 48:W03521.

Li, S. and Jain, A. K. (2009). Forward-backward algorithm. In Li, S. and Jain, A. K., editors, *Encyclopedia of Biometrics*, chapter 10, pages 266–290. Springer, Boston.

Lister, D., Osborn, T., Jones, P., and Darch, G. (2018). An assessment of extreme drought rainfall sequences in stochastic weather generator output for the greater Anglian region. https://sites.uea.ac.uk/documents/421974/1301877/ENV-CRU-RP23-Reduced-2019.pdf/550df0f0-8524-30ca-d4e5-d70f8162f48b.

Lowe, J. A. et al. (2018). UKCP18 science overview report. Technical report.

Monma, C., Paterson, M., Suri, S., and Yao, F. (1990). Computing euclidean maximum spanning trees. *Algoritmica*, 5:407–419.

Naveau, P., Huser, R., Ribereau, P., and Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4):2753–2769.

Nelson, R. B. (2006). *An Introduction to copulas: 2nd Edition*. Springer, New York.

Oriani, F., Mehrotra, R., Mariethoz, G., Straubhaar, J., Sharma, A., and Renard, P. (2018). Simulating rainfall time-series: how to account for statistical variability at multiple scales? *Stochastic Environmental Research and Risk Assessment*, 32:321–340.

Ossó, A., Sutton, R., Shaffrey, L., and Dong, B. (2018). Observational evidence of European summer weather patterns predictable from spring. *Proceedings of the National Academy of Sciences*, 115:59–63.

Perry, M. and Hollis, D. (2005). The generation of monthly gridded datasets for a range of climatic variables over the UK. *International Journal of Climatology*, 25:1041–1054.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11:5018–5051.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, 108:4407.

Serinaldi, F. and Kilsby, C. G. (2012). A modular class of multisite monthly rainfall generators for water resource management and impact studies. *Journal of Hydrology*, 464–465:528–540.

Serinaldi, F. and Kilsby, C. G. (2014a). Rainfall extremes: Toward reconciliation after the battle of distributions. *Water Resources Research*, 50:336–352.

Serinaldi, F. and Kilsby, C. G. (2014b). Simulating daily rainfall fields over large areas for collective risk estimation. *Journal of Hydrology*, 512:285–302.

Stoner, O. and Economou, T. (2020a). An advanced hidden markov model for hourly rainfall time series. *Computational Statistics and Data Analysis*, 152:107045.

Stoner, O. and Economou, T. (2020b). A coupled hidden markov model for daily rainfall at multiple sites. Proceedings of the 35th International Workshop on Statistical Modelling.

Sutton, R. and Dong, B. (2012). Atlantic Ocean influence on a shift in European climate in the 1990s. *Nature Geoscience*, 5:788–792.

Sutton, R. T. and Hodson, D. L. R. (2005). Atlantic Ocean forcing of North American and European summer climate. *Science*, 309(5731):115–118.

Timonina, A., Hochrainer-Stigler, S., Pflug, G., Jongman, B., and Rojas, R. (2015). Structured coupling of probability loss distributions: Assessing joint flood risk in multiple river basins. *Risk Analysis*, 35:2102–2119.

van Erp, S., Oberski, D. L., and Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50.

Verdin, A., Rajagopalan, B., Kleiber, W., , Podesta, G., and Bert, F. (2018). A conditional stochastic weather generator for seasonal to multi-decadal simulations. *Journal of Hydrology*, 556:835–846.

Verdin, A., Rajagopalan, B., Kleiber, W., and Katz, R. W. (2015). Coupled stochastic weather generation using spatial and generalized linear models. *Stochastic Environmental Research and Risk Assessment*, 29:347–356.

Verdin, A., Rajagopalan, B., Kleiber, W., Podesta, G., and Bert, F. (2019). BayGEN: A bayesian space-time stochastic weather generator. *Water Resources Research*, 55(4):2900–2915.

Wadsworth, J. L. and Tawn, J. A. (2012). Dependence modelling for spatial extremes. *Biometrika*, 99:253–272.

Water UK (2016). Water resources long term planning framework (2015-2065), Atkins, Mott MacDonald, Nera, HR Wallingford and Oxford University, for Water UK. Technical report.

Wilby, R. L., O'Hare, G., and Barnsley, N. (1997). The north atlantic oscillation and british isles climate variability, 1865-1996. *Weather*, 52:266–276.

Wilks, D. S. (2008). High-resolution spatial interpolation of weather generator parameters using local weighted regressions. *Agricultural and Forest Meteorology*, 148(1):111–120.

Wilks, D. S. (2009). A gridded multisite weather generator and synchronization to observed weather data. *Water Resources Research*, 45:W10419.

Wilks, D. S. and Wilby, R. L. (1999). The weather generation game: a review of stochastic weather models. *Progress in Physical Geography: Earth and Environment*, 23(3):329–357.

Yevjevich, V. (1968). Misconceptions in hydrology and their consequences. *Water Resources Research*, 4(2):225–232.

Acknowledgements

The authors wish to thank: Dan Hollis for support with running the NCIC interpolation scheme; Jill Dixon for the extraction of validation data; Elizabeth Brock for help in understanding previous stochastic rainfall work in the Greater Anglian region; Emily Wallace, Jessica Standen and Malcolm Lee for scientific review; and the peer reviewers whose comments have helped to improve and clarify this manuscript. This work was funded by Anglian Water.

Data statement

The HadUK-Grid rainfall data can be found at https://catalogue.ceda.ac.uk/uuid/4dc8450d889a491ebb20e724debe2dfb (Hollis et al., 2019). Data used to calculate MSLP, EMSLP is at https://www.metoffice.gov.uk/hadobs/emslp/ (Ansell et al., 2006), ERA5 at https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5, and HadISST data can be accessed at https://www.metoffice.gov.uk/hadobs/hadisst/ (Rayner et al., 2003).

The code and data developed within this study cannot be shared due to ownership of intellectual property.

Author contributions statement

L.D.: Conceptualization, Methodology, Software, Validation, Formal Analysis, Writing, Visualization.

J.O.: Project administration; Methodology, Software, Validation, Formal Analysis, Writing, Visualization.

T.E.: Conceptualization, Supervision.

G.D.: Resources; Supervision.

O.S: Software, Supervision.