Investigation of probabilistic, deterministic and hybrid models for improved characterisation of hydrological extremes across varying temporal scales

Stelian Curceac, DCE, MSc

Lancaster Environment Centre

Lancaster University

Submitted for the degree of Doctor of Philosophy

January 2022

Abstract

Peak water flow events increase the risk of flooding which can have severe negative impacts on human lives and ecosystem services. Moreover, high water run-off from agricultural land increases sediment and nutrient losses that can result in soil degradation and water course pollution. In this thesis, peak flow events were modelled using statistical and machine learning approaches, process-based models (PBM) and a combination of the two. In the first thesis study, high-flow data measured over a period of 6 years (2012-2018) at the North Wyke Farm Platform, an agricultural research facility in south-west England, were characterised by the Generalised Pareto distribution (GPD). Based on the analysis of the effects of GPD parameter estimators, sample size and different temporal resolutions (15 mins, hourly, 6 hourly and daily), an automated threshold selection method based on stability plots was proposed to define peak flow events. This method was evaluated using diagnostic indices and Quantile-Quantile plots and its advantages were demonstrated. For the second study, an existing PBM (SPACSYS) was used to simulate flow at four temporal resolutions: (i) the daily resolution which is the resolution it was first developed to run at, (ii) 15 mins, (iii) hourly and (iv) 6 hourly. The simulated flow was compared to the measured values at each of the four data resolutions and also via an aggregation to the coarsest daily scale. Model performance graphics and calculated accuracy statistics showed that simulating at finer resolutions and then upscaling to the daily scale provided a more accurate representation of the number and magnitude of peak flow events. The third study, focused on improving daily PBM simulations of peak flow events by using a hybrid modelling framework where the same PBM was combined with a statistical model that stems from Extreme Value Theory (Conditional

Extreme Model) and a data-driven machine learning model (Extreme Learning Machine). Assessed by goodness-of-fit indices, such as the mean absolute error (MAE), the normalized root mean square error (NRMSE), the percentage BIAS (PBIAS), the Nash-Sutcliffe efficiency (NSE), the index of agreement (d) and the Kling-Gupta Efficiency (KGE), the proposed hybrid approach was better able to capture the dynamics of the peak flow events and increase the accuracy of the predictions. For the first three studies, all methods were largely evaluated from a prediction viewpoint using error and agreement indices described above. The fourth and final thesis study, explored the use of variograms and wavelets to assess the performance of the proposed models in terms of capturing measured flow variation at different temporal scales, and in the context of peak flow events. It built on the findings from the previous studies as the hybrid model was also applied on hourly aggregated to daily PBM simulations. The use of soil moisture as a covariate was also investigated. A change point analysis found that the magnitude of the local wavelet variance was related to the frequency of peak flow events and the days before they occurred. As a whole, this thesis provides clear advances, via a series of linked studies for improved identification and characterisation of modelled peak water flows across different temporal scales.

Declaration

Except where reference is made to other sources, I declare that the work presented in this thesis is the authors own and has not been submitted to any other institutions for any degree or qualifications. Collaborations with other researchers and publications are properly acknowledged.

Stelian Curceac

Lancaster University, January 2022.

Statement of authorship

Part of this thesis has previously been published in a peer-reviewed journal in collaboration with other authors. All four co-authors are part of my PhD supervisory team. Details of the publication, authorships and author contributions are listed below:

Chapter 2 - Paper 1: Curceac, S., Atkinson. P. M., Milne, A., Wu, L. and Harris, P. (2020). An evaluation of automated GPD threshold selection methods for hydrological extremes across different scales. *Journal of Hydrology*, 585, 124845.

Chapter 3 - Paper 2: Wu, L., Curceac, S., Atkinson, P. M., Milne, A. and Harris, P. (2020). A case study on the effects of data temporal resolution on the simulation of water flux extremes using a process-based model at the grassland field scale. *Agricultural Water Management,* 255, 107049*.*

Chapter 4 - Paper 3: Curceac, S., Atkinson. P. M., Milne, A., Wu, L. and Harris, P. (2020). Adjusting for conditional bias in process model simulations of hydrological extremes: an experiment using the North Wyke Farm Platform. *Frontiers in Artificial Intelligence, AI in Food, Agriculture and Water, Research Topic: Machine Learning for Water Resources, 3:565859*

Chapter 5 - Paper 4: Curceac, S., Milne, A., Atkinson, P.M., Wu, L., Harris, P. (2020) Elucidating the performance of hybrid models for predicting extreme water flow events through variography and wavelet analyses, *Journal of Hydrology,* 598, 126442*.*

Acknowledgements

This work was funded by Rothamsted Research and Lancaster Environment Centre, the BBSRC Institute Strategic Programme (ISP) grant, "Soils to Nutrition" (S2N) grant numbers BBS/E/C/000I0320, BBS/E/C/000I0330 and the BBSRC National Capability grant for the North Wyke Farm Platform grant number BBS/E/C/000J0100. The thesis datasets are open and freely available from: https://www.rothamsted.ac.uk/north-wyke-farm-platform and the SPACSYS model can be found here: https://www.rothamsted.ac.uk/rothamsted-spacsys-model. R software (R Core Team, 2020, https://cran.r-project.org/) was used throughout the thesis. For the implementation of the hybrid models in Chapters 4 and 5, the CEM is from the texmex R package (Southworth et al., 2020) and the ELM from the elmNNRcpp R package (Mouselimis and Gosso, 2020). Model performance indices were calculated by using functions in the hydroGOF R package (Zambrano-Bigiarini, 2017). The variogram analysis of Chapter 5 was conducted in GENSTAT (VSN International, 2019).

I would like to thank all my supervisors, Paul Harris, Peter Atkinson, Alice Milne and Lianhai Wu for all their help and support. I am really grateful to them for their time, patience and encouragement. I would also like to extend my deep thanks to everyone at Lancaster University and Rothamsted Research who helped in any way for this accomplishment. I am also deeply thankful to my family for their faith in me. Of course thanks Pauline for standing next to me until the very last minute before submission and making this journey so much more exciting.

Table of contents

List of figures

List of tables

Glossary of terms

AD	Anderson-Darling
AM	Annual Maxima
ANN	Artificial Neural Network
ATS	Automated Threshold Stability
BBSRC	Biotechnology and Biological Sciences Research Council
C	Carbon
CDF	Cumulative Distribution Function
CEM	Conditional Extreme Model
CvM	Cramer-von Mises
d	index of agreement
ELM	Extreme Learning Machine
EVT	Extreme Value Theory
FFA	Flood Frequency Analysis
GEV	Generalized Extreme Value
GHG	Greenhouse Gas
GP	Generalized Pareto
GPD	Generalised Pareto Distribution
HN	Hidden Neuron
H2D	Hourly to Daily
ISP	Institute Strategic Programme
KGE	Kling-Gupta Efficiency
KS	Kolmogorov-Smirnov

LME	Likelihood Moment Estimator
MAE	Mean Absolute Error
MGF	Maximum Goodness-of-Fit
MJO	Madden-Julian Oscillation
MLE	Maximum Likelihood Estimator
MODWT	Maximum Overlap Discrete Wavelet Transform
MOM	Method of Moments
MPLE	Maximum Penalized Likelihood Estimator
MRA	Multi-Resolution Analysis
MRL	Mean Residual Life
MSE	Mean Square Error
N	Nitrogen
NAO	North Atlantic Oscillation
NRMSE	Normalized Root Mean Square Error (NRMSE
NSE	Nash-Sutcliffe Efficiency
NWFP	North Wyke Farm Platform
P	Pearson’s chi-square
PBIAS	Percentage Bias
PBM	Process-Based Model
PCA	Principal Components Analysis
Pick	Pickland’s estimator
POT	Peaks Over Threshold
PWM	Probability Weighted Moment
Q-Q	Quantile-Quantile

RD	Relative Index of Agreement
SE	Square Error
SF	Shapiro-Francia
SM	Soil Moisture
S2N	Soils to Nutrition

1. Introduction

1.1 Background and extreme events in hydrology

For extremes in hydrology, peak water flow events increase the possibility of flooding which is one of the most common natural disasters. Globally, over 1 billion people are estimated to be potentially affected by floods which can cause immense damage to properties and loss of thousands of lives annually (Guha-Sapir et al., 2016; Jonkman and Vrijling, 2008). In the UK, more than 5 million people in 2.4 million properties are at risk of flooding (Environment Agency, 2009) and the estimated yearly cost of damages caused by floods is over £1 billion (Collet et al., 2017). The changing patterns of rainfall events may intensify the existing risks posed by flooding, as the magnitude and frequency of floods is possible to increase as a result of a changing climate (Bates et al., 2008; Field et al., 2012; Kundzewicz et al., 2007). This is likely to increase flooding in farmland areas and put at risk current agricultural capabilities (Brown et al., 2016). Other significant impacts are soil degradation and water course contamination caused by increased nutrient and sediment losses during high runoff (Bouraoui et al., 2004). Areas with steep and unstable slopes are more vulnerable to landslides during heavy precipitation that results in high runoff (Clarke and Rendell, 2006). Extreme hydrological events can severely alter the dynamics of ecological communities and push ecosystems beyond the threshold of normal disturbance resulting in irreversible impacts (e.g. Thibault and Brown, 2008).

Interestingly, recent studies on historical streamflow data do not suggest an increase in extremes and the projected increase of precipitation extremes will actually not be associated with an increase in high streamflow and flooding on account of reduced soil

moisture due to temperature rise and reduced storm durations (Do et al., 2017; Hall et al., 2014; Hodgkins et al., 2017; Wasko et al., 2019). Furthermore, extreme events can have positive impacts such as the reproduction of ecosystems through floods, disease control, population control for ecosystem balance or general ecosystem health. Therefore, regardless of the (positive or negative) direction of the consequences of flooding, accurate and reliable modelling and forecasting of extreme flow events is crucial for water resources planning, impact assessment and implementing measures to mitigate their effects and so protect lives, properties and services.

Defining extreme events is complex and there is no one universal definition of extremes. Various definitions and approaches can be found in the scientific literature. According to IPCC (2012), an extreme climate or weather event is “the occurrence of a value of a weather or climate variable above (or below) a threshold value near the upper (or lower) ends of the range of observed values of the variable”. A threshold can be defined as a relative or an absolute value. In the first case, it depends on the observed values of a specific variable. For example, Hansen et al. (2012) define extreme temperature as values that are higher than three standard deviations above the mean. In other studies, percentiles or probabilities of occurrence of 1, 5 or 10 %, as well as values that have a large return period (e.g. 50-year event) are preferred. Absolute thresholds usually refer to conditions that are critical for a certain activity (e.g. floods causing damages or suspended functionality of economic sectors such as production, transportation and communication). Other factors that must be taken into consideration when defining extremes are the duration and intensity, the affected area, timing, frequency, continuity-discontinuity and

pre-conditioning (e.g. a heavy rainfall is more likely to cause flooding if the river flow is already higher than usual).

1.2 Extreme Value Theory

A core approach used for modelling hydrometeorological variables is statistical and data-driven models, where an abundance of methods can be found in the literature (Solomatine and Ostfeld, 2008; Elshorbagy et al., 2010). Extreme Value Theory (EVT) is the branch of statistics which is used to describe the stochastic behaviour of rare events. In the context of hydrology, this refers to the analysis of the relationship between flood magnitude and its corresponding frequency of occurrence, where Flood Frequency Analysis (FFA) is implemented. By definition, rare events, such as peak discharge, are in the tail of the distribution function. As most of the observations are gathered towards the centre of the distribution, only a limited number is in the tails.

There are two parametric probability distributions that are used to model the tails of distributions, depending on how the extremes are defined. The first one is the block (usually annual) maxima (AM) where the dataset is split into adjacent blocks of the same size and the maximum value in each block is taken. These independent and identically distributed (iid) variables asymptotically follow a Generalized Extreme Value (GEV) distribution (Fisher and Tippett, 1928; Jenkinson, 1955; Coles, 2001). The second one is the Peaks Over Threshold (POT), where the values that exceed a high enough threshold (which defines the extremes) are modelled by a Generalised Pareto Distribution (GPD) (Pickands, 1975).

These two families of distributions have fundamental differences and theoretical links (Langousis et al., 2016). Both have 3 parameters that need to be estimated, namely the

location, scale and shape parameters. The shape parameter is probably the most important and the most difficult to estimate. It is believed that the shape parameter describes an inherent feature of the process and is not affected by changes in the scale of the observations. For example, peak flows generally exhibit positive shape characteristics even if they have differences of order of magnitude, which is depicted in the scale parameter. The value of the shape parameter indicates if the large values increase faster and are asymptotically infinite (positive) or increase more gradually and are upper-bounded (negative). A process described by a zero valued shape parameter behaves exponentially (Figure 1-1). The values of the parameters can vary significantly according to the estimators and the sample size (Engeland et al., 2004). The size of the sample depends on the length of the series and the size of the blocks for AM and the threshold for POT. A GEV approach is best applied to annual maxima and consequently, a long series of data are required. This restriction does not apply for GPD as all the peaks above a threshold are considered. However, this criterion should not be over-stretched and the peaks are required to be independent. The selection of the threshold is also a grey area in the literature although many different techniques have been proposed (see Scarrott and MacDonald, 2012 and Langousis et al., 2016 for extensive reviews). A high threshold ensures that the excesses above it follow a GPD but can lead to reduced sample size which increases the uncertainty of the estimates. A smaller threshold increases the sample size but also the bias of the estimates as the empirical distribution deviates from a perfect GPD model (Scarrott and MacDonald, 2012).

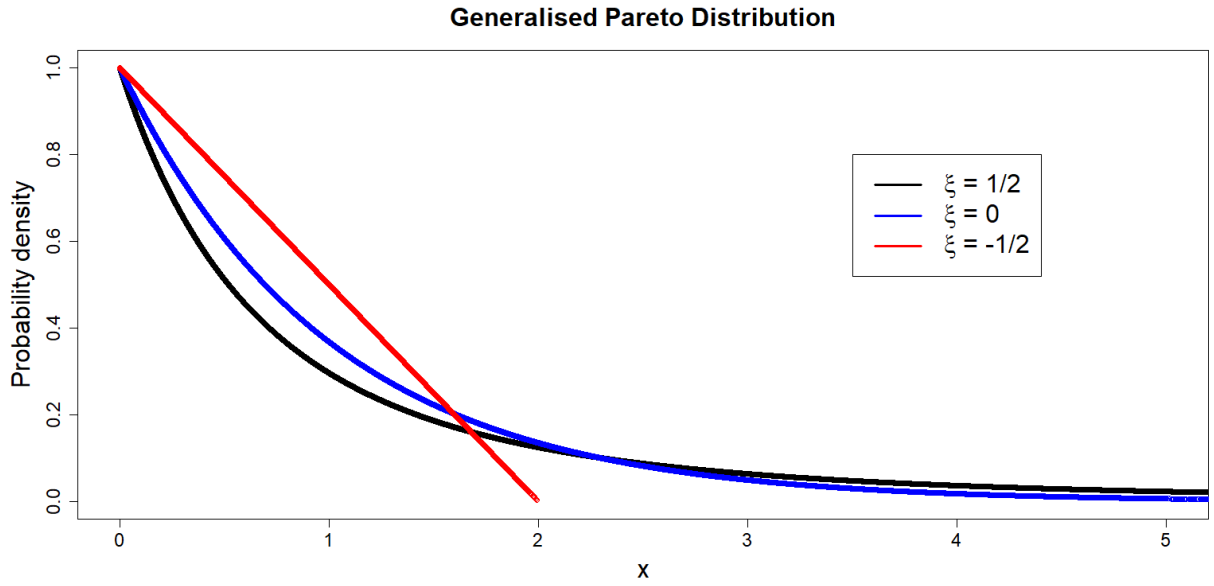


Figure 1-1: PDF of the GPD for different values of the shape parameter ξ

1.3 Process-based modelling

Modelling and forecasting of water flow or streamflow has been traditionally performed by physically based models. Similar modelling approaches are also known in the literature as process-based or deterministic or lumped models. A process-based model (referred to here as a PBM) is the mathematical representation of one or more natural processes that take place within a defined system. It can vary from a simple form consisting of one ordinary or differential equation with a few parameters to very complex multiparametric models that contain a large number of equations which describe various processes. PBMs sometimes suffer from over-parameterization or excessive complexity and can be difficult to use. However, they have been frequently proven to be very useful tools to help scientists (i) understand complex interactions within natural systems and interpret experimental results (Peck, 2004), (ii) test multiple scenarios which are infeasible to evaluate by experimentation (Asseng et al., 2013) (iii) support decisions (Uusitalo et al., 2015) and (iv) predict the outcome of certain actions (Beechie et al., 2010).

The incorporation of different physical hydrological processes in a PBM started with the works of Crawford and Linsley (1966) and Freeze and Harlan (1969). In the recent 50+ years, a considerable amount of research has been conducted in their development making them a very useful tool for hydrologists. The collaboration with other disciplines of geosciences, such as ecohydrology and geomorphology (Fatichi et al., 2016), has been necessary for this progress in order to manage, capture and model the various processes that take place at a watershed scale. They also have a high degree of flexibility for allowing changes in the parameterisation in order to incorporate alterations in watershed characteristics, land cover changes and climate change for better short-term and/or long-term forecasts.

The key hydrological processes described in most process-based models are precipitation, surface runoff, evapotranspiration, infiltration, lateral flow, percolation, soil erosion and sediment losses. Not all of these main processes are included in all the models and some models have other processes incorporated, depending on the objective of the model, as well as data availability. The two main groups of process-based models for hydrological modelling are conceptual and physical models. In physical models, various hydrological processes are represented by equations of mass, momentum and energy conservation. They have the ability to model the spatial variability of land use, slope, soil characteristics and climate conditions within the watershed semi or fully distributed in nature. For this reason, a large number of parameters describing the physical characteristics of the catchment need to be calibrated and consequently the data requirement is also large. As the parameters have a physical interpretation, these types of models have been shown to provide accurate hydrologic forecasts (Abbott et al., 1986; Arnold et al., 1998; Carpenter

and Georgakakos, 2006; El Hassan et al., 2013; Jaiswal et al., 2020). The conceptual hydrological models use semi-empirical equations and consists of a number of interconnected storages which represent the physical elements of a catchment which deplete and recharge. Unlike physical models, conceptual models do not represent spatial variation , instead the model parameters are averaged over the whole catchment. As a result, they require far fewer parameters than physical models. However more data is generally needed for their calibration and they are more sensitive to the input data (Schumann, 1993; Donnelly-Makowecki and Moore, 1999).

The most common climatological input data for hydrological PBMs are precipitation and air temperature. Other inputs include soil characteristics, topography, vegetation, hydrogeology and other physical parameters. Where possible, measured data is used as input. In cases where observational records are not available, predicted input data can be obtained, for example from regional climate models (Akhtar et al., 2009), where downscaling techniques are commonly applied for the data to be at the appropriate resolution (Chen et al., 2012). However, the uncertainty that forecasted data introduces needs to be accounted for, as it can have significant impact on models' performance (Kobold and Sušelj, 2005; Arnaud et al., 2011). Accounting for the uncertainty in the parameter estimates, input data, model structure and calibration/validation data has been shown to reduce the bias in the forecasts and provide more realistic predictions (Yen et al., 2014).

The estimation of the potential evapotranspiration (PET), a significant component of hydrological modelling, is another source of uncertainty. However, different estimates of PET can provide similar runoff simulated values (Bai et al., 2016). Similarly, differences in

the simulations of various other highly-nonlinear hydrological processes can frequently lead to the same result. The reason for this is that the parameters are estimated by optimisation algorithms based on objective functions and a change in one parameter is compensated by changes in the other parameters, which can result in non-feasible parameter vectors that do not have any physical meaning (Bardossy and Singh, 2008; Arnaud et al., 2011). In the scientific literature, this phenomenon is called equifinality (Beven 1975, 1993), where the same or similar outcome can be achieved by different pathways, and various facets of equifinality have received attention in the recent years (Khatami et al., 2019).

PBM performance in simulating streamflow depends on their structure, site or catchment characteristics (e.g. hydrological regime), seasonality and climatology. For example, hydrological models tend to be more accurate in wetter catchments during winter (Lidén and Harlin, 2000; McMillan et al., 2016). A common characteristic is the under-prediction of flow extremes (Lane et al., 2019; Wijayarathne and Coulibaly, 2020). Most frequently, they exhibit a smoothing effect which results in over-prediction of low flows and under-prediction of high flows (McMillan et al., 2016; Newman et al., 2015). This behaviour is caused by systematic errors (conditional bias), where the structure and parameters of the model generalise the outputs, which is an effect similar in nature to that of having a support that is larger than ideal. Models' performance can also suffer random errors, which are due to uncertainties of the estimated parameters and the inadequate quality of the input data.

1.4 Hybrid modelling

As mentioned above, PBMs are the mathematical representation of the laws of physics and their implementation is deterministic. In contrast, statistical models do not depict the laws of physics but describe the observed patterns of the data and extract information from it. A common perception is that these two discrete approaches belong to totally different and unbridgeable schools of thought. Indeed, by their nature, the two methods have their own specific characteristics and are associated with a distinct spirit and way of thinking. However, according to Berliner (2003), these two models are the endpoints of the same spectrum since they are interrelated to some extent. For example, observed datasets are necessary for the development and assessment of the physical models and for the estimation of their parameters and boundary and initial conditions. Respectively, the development of statistical models is based on the fundamental principles of sciences such as mathematics and physics. The combination of statistical and physical reasoning started in the 1960s with the works of R. E. Kalman which resulted in the commonly used Kalman filter (Kalman, 1960). Since then, various approaches have been proposed for the combination (or hybridisation) of the two modelling methods, with an increasing trend in development in recent years (Toth et al., 1999; Cloke and Pappenberger, 2009; Bogner et al., 2017; Papacharalampous et al., 2019).

The growing computing capability allows physical models to be run in parallel for a range of parameters and thus provide multiple outputs (ensembles) which can then be analysed using statistics in a post-processing setting (Cloke and Pappenberger, 2009; Li et al., 2017). This can provide valuable information regarding the model's structure and which aspects of it to target, in terms of improving performance and expressing the uncertainty of the

predictions in terms of confidence intervals or predicted distributions rather than a single data point. For this type of analysis, Bayesian methods are most commonly applied (Berliner, 2003; Raftery et al., 2005; Bradley et al., 2015). Other post-processing methods that can be found in the literature are stochastic data-driven methods on wavelet decomposed series (Quilty et al., 2019), extended logistic regression (Roulin and Vannitsem, 2011), quantile regression (López López et al., 2014), bias correction (Li et al., 2019) and nearest neighbor resampling for uncertainty estimation (Sikorska et al., 2015), to name a few.

One of the first research studies demonstrating an increase of accuracy in real-time flood forecasting by combining deterministic and stochastic (Box and Jenkins, 1976) models was performed by (Toth et al., 1999). There has also been a recent increase in the development and testing of various post-processing techniques in the field of hydrology (Schaake et al., 2006; Brown and Seo, 2010; Zhao et al., 2011; Hemri et al., 2015). The performance of the different proposed approaches has been found to vary significantly depending on the level of the streamflow (Bogner et al., 2016; Papacharalampous et al., 2019), as well as catchment characteristics and hydrological conditions (Dogulu et al., 2015). Forecasts from hydrological models can also be combined with more than one statistical or data-driven models (Bogner et al., 2017). Except for the combination of physically-based and statistical methods, hybrid modelling approaches can also include the combination of classical statistical methods with more data-driven, machine-learning methods, such as artificial neural networks (ANNs) (Yaseen et al., 2016; Chen et al., 2018; Zhou et al., 2018), discrete wavelet transforms and support vector machines (Kisi and Cimen, 2011), and the coupling of ANNs with autoregressive techniques (Fathian et al., 2019).

1.5 Scale issues

An important issue that affects the performance of statistical, process-based and hybrid models, regarding total runoff and peak flow, is the time step or more generally the time scale. This is similarly true for the spatial scale in a spatial analysis framework (Schaake et al. 1996; Haddeland et al., 2002; 2006), where spatial discretisation in a cell grid (distributed models) has been found to have better performance compared to models where sub-basins are the unit of analysis (semi-distributed) (see Caldeira et al. 2019). The time step usually depends on the objective for which the model is used. For example, hourly or even minute simulations are required for flood forecasting whereas simulated flow at daily or monthly resolution are appropriate for reservoir and water supply management. Since the 1990s, the increase in availability of data at hourly and even sub-hourly resolutions has allowed the development of plethora of hydrological models running at sub daily time steps (Ficchì et al., 2016). Studies have shown that the accuracy of the higher resolution / decreased time step simulations generally increases, and is less sensitive to the model's parameters or the spatial resolution (Jeong et al., 2010; Choi et al., 2018; Huang et al., 2019). Running models at sub-daily time steps can also significantly increase the accuracy of flood forecasting (Kobold and Brilly, 2006). However, for PBMs, this requires the input variables, such as precipitation and temperature, to be at a corresponding high resolution, which can introduce additional uncertainty due to data quality issues (Yu et al., 1997). When data at sub-daily resolution is not available, various temporal disaggregation methods have been proposed and reported reliable (Pui et al., 2012; Bárdossy and Pegram, 2016; 2017; Breinl and Di Baldassarre, 2019). Using hydrological models at finer time scales is usually straightforward but recalibration of

algorithms and parameters is likely to be necessary (Jeong et al., 2010; Kavetski et al., 2011).

1.6 Model assessment

In any given model development exercise, the newly-proposed model should be validated and given context against a range of existing models using either separate training and validation datasets or just a single dataset with a resampling technique, such as cross-validation (Borra and Di Ciaccio, 2010). . For each model, a range of accuracy, precision and bias diagnostics can be found based on the resultant errors (i.e. measured minus predicted) (Smith et al., 1997). Such global diagnostics are often supplemented through measured versus predicted scatterplots with the 45 degree line through the origin drawn to see model deviations from the ideal fit. There is also scope for local diagnostics, to see at which specific period on the flow hydrograph, the given model performs strongly or weakly (see, Harris et al., 2013; Tsutsumida et al., 2019 in a spatial context). In the context of peak flow prediction, a threshold based on the measured data can be set to determine if modelled peaks similarly exceed this threshold. Incidences of correct peak flow predictions, false negatives (prediction does not exceed threshold when measured flow does), false positives (prediction exceeds threshold when measured flow does not). The kappa statistic (κ)can be computed:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o represents the actual observed agreement and p_e represents chance agreement. This statistic provides a measure of agreement beyond the level of agreement expected by chance alone. Again, kappa provides a global assessment, and analogous to the localised

diagnostics, above, a local kappa can be found to provide a more detailed analysis (see Comber et al., 2017 in a spatial context). Further, it is not only possible to assess model prediction accuracy, it is also possible through a wavelet analysis to provide a local assessment of the modelled flow in terms of it accurately capturing measured flow variation across different temporal scales (Rust et al., 2014).

1.7 Study site and data

To demonstrate the methodological advances presented in this thesis, flow discharge data measured at the North Wyke Farm Platform (NWFP) was used as case study data. The NWFP is a farm-scale experiment established in 2010 in the southwest of England (50°46'10"N, 3°54'05"W) to support research into sustainable grassland livestock systems (Orr et al., 2016). The platform comprises three independent small farms, each 21 ha in size. Each farm is divided into five sub-catchments, with some sub-catchments consisting of more than one field. The platform monitors routinely water run-off and water chemistry in each of the 15 sub-catchments, together with other primary data collections (e.g. greenhouse gas emissions) so that each farming system can be evaluated according to its level of sustainability (Takahashi et al., 2018). For the period 1985-2015, the average annual temperature at North Wyke ranges from 6.8 to 13.4 °C and the average annual rainfall is 1033 mm.

The platform has an altitude range of 120–180 m above sea level. Soil texture consists of a slightly stony clay loam topsoil (about 36% clay) above a mottled stony clay (about 60% clay). The subsoil is impermeable to water and during rain events most of the excess water moves by surface and subsurface lateral flow towards the drainage system described below. North Wyke is underlain by the Carboniferous Crackington Formation, a part of

what geologists have called ‘Cu1m Measures’ (Harrod and Hogan, 2008). The Crackington Formation comprises clay shales (locally known as ‘shillot’) with thin subsidiary sandstone bands. The shales which may be somewhat cleaved, are dark grey or black, but weather pale brown or buff. When waterlogged they break down readily to form clay, the clay minerals being predominantly illitic. Sandstone bands in the Crackington Formation probably comprise about a quarter of the sequence but are rarely thicker than 30-40 cm. As a whole, the Carboniferous dips to the north and is affected by folds with east-west axes. The restricted number of local exposures, mainly in the riverbed, indicate steep, near vertical dips, with some overturning. As over much of Devon, in situ rocks are largely mantled with Head of varying (0.5-3 m) thickness. This is rock waste of local origin resulting from protracted frost working and solifluction during the Pleistocene. A very small igneous dyke runs east north-east to west south-west across the ridge at North Wyke itself, the outcrop being partly picked out by quarried ground. The rock is altered by weathering from its original state (lamprophyre, a medium grained, intermediate igneous rock), and is not of account as a soil parent material here.

Each of the 15 sub-catchments (Figure 1-2)are hydrologically isolated through a combination of topography and a network of French drains totalling 9.2 km in length and constructed by digging 800-mm deep trenches, lining them with damp proof membrane, and placing a perforated plastic drainage pipe centrally in the trench bed. The width of the drains was dependent on the drainage pipe diameter + 100mm each side and to facilitate this, eight different digger bucket sizes were fabricated. The trenches were backfilled using 5056 tonnes in total of 20 – 50 mm clean granite stone. All the flumes receive water supplied by 2 branches of the drains and where these join in a confluence pit, puddled clay

was placed around the pipe to ensure the drainage water is always captured. The water is then channelled via concrete piping and a sampling pit into the flume. Where required, the experimental areas have been protected on the upslope boundaries by open ditches and sealed pipes to prevent ingress of external groundwater and surface runoff from adjacent land. Each flume is supplied with mains electric power and a fibre optic cable based data telemetry system which totals over 5 km in length.

Thus each of the sub-catchments drains to a single monitoring station (at the flume), where the quantity of discharge from each sub-catchment is measured through a combination of primary and secondary flow devices. The primary devices are H Type flumes [TRACOM Inc., Georgia, USA] with capacity designed for a 1 in 50-year storm event. Flumes are fixed engineered structures that intercept and channel free-flowing liquids in such a way that flow rate can be determined by a known relationship (rating curve) between the height of the liquid at a single specific location in the flume and its flow rate. The specific design of the H flume facilitates the accurate measurement of both low and high flows and is relatively self-cleaning since it allows the ready passage of sediment and particulate matter [ISCO open channel flow measurement handbook, 2008]. The choice in size of the flumes installed on the NWFP was determined by size of the catchment they are servicing and are 450mm, 600mm and 750mm. Pressure level sensors [OTT hydromet, Loveland, CO., USA] are the secondary devices that are used to measure the depth of water by means of an integrated controller and ceramic pressure-measuring cell. The output data are converted to flow (L s^{-1}) externally using water height specific formulae (**Error! Reference source not found.**). Each catchment site has a cabin or flume laboratory which houses telemetry devices, pumping equipment, and a by-pass flow cell which contains sensors to measure

various water quality parameters. The flow is generated only from rainfall as the fields are not irrigated. In this research, measured and model simulated flow data are used from sub-catchment 3 (consisting of two fields - Poor Field and Ware Park) and sub-catchment 6 (Golden Rove), for the periods 2012 to 2018 and 2013 to 2016, respectively (Figure 1-3).

Table 1-1: Formulae for conversion of water height to discharge rate for different sized flumes.

Catchment Number	Flume Size (mm)	Formulae (H in metres)
1, 7, 10, 11, 12, 13, 14, 15	450	L^{-s}=-0.00396-0.07232*$H^{0.5}$)+(79.89379*$H^{1.5}$)+(900.3765*$H^{2.5}$)
2, 3, 5, 6, 8, 9	600	L^{-s}=-0.022285-(0.55496*$H^{0.5}$)+(125.5276*$H^{1.5}$)+(939.5717*$H^{2.5}$)
4	750	L^{-s}=-0.042447-(0.90725*$H^{0.5}$)+(108.676*$H^{1.5}$)+(937.5943*$H^{2.5}$)

Each of the 15 catchments has a soil moisture station (SMS) sited at a central location (Figure 1-2), consisting of a remote telemetry units (RTU), a combination soil moisture and soil temperature probe and a rain gauge (RG) [Adcon, Austria]. The soil moisture probe measures soil moisture through capacitance at depths of 10cm, 20cm and 30cm, and soil temperature at 15cm. However, only soil moisture data at 10cm are available on the data portal as data at the lower depths were deemed unreliable for this soil series. The direct connection to the RTU is via an SDI 12 interface and the raw data is converted to soil moisture using a lookup table developed from testing the sensor output in blocks of North Wyke soil at a range of conditions. Data from the tipping bucket rain gauge are collected by the RTUs integrated pulse counter at a resolution of 0.2mm per tip.

Two sets of instruments are co-located at the meteorological station. More specifically, these are an Official UK Meteorological Office equipment and a NWFP dedicated equipment, with data collected since 29th April 2013. The following meteorological

variables are collected by the site-specific instruments: precipitation (mm) (installed in Nov 2011), air temperature (°C), relative humidity (%), wind speed (km/h), wind direction (in degrees) and solar radiation (W/m2, installed in May 2014). NWFP meteorological data are collected at 15 min intervals. The tipping bucket rain gauge was phased out and replaced by a more accurate, Pluvio weighing RG installed in 12th April 2015 which can provide precipitation (mm) data at 1 min intervals (but currently not exported as such) and data for this is available from August 2013.

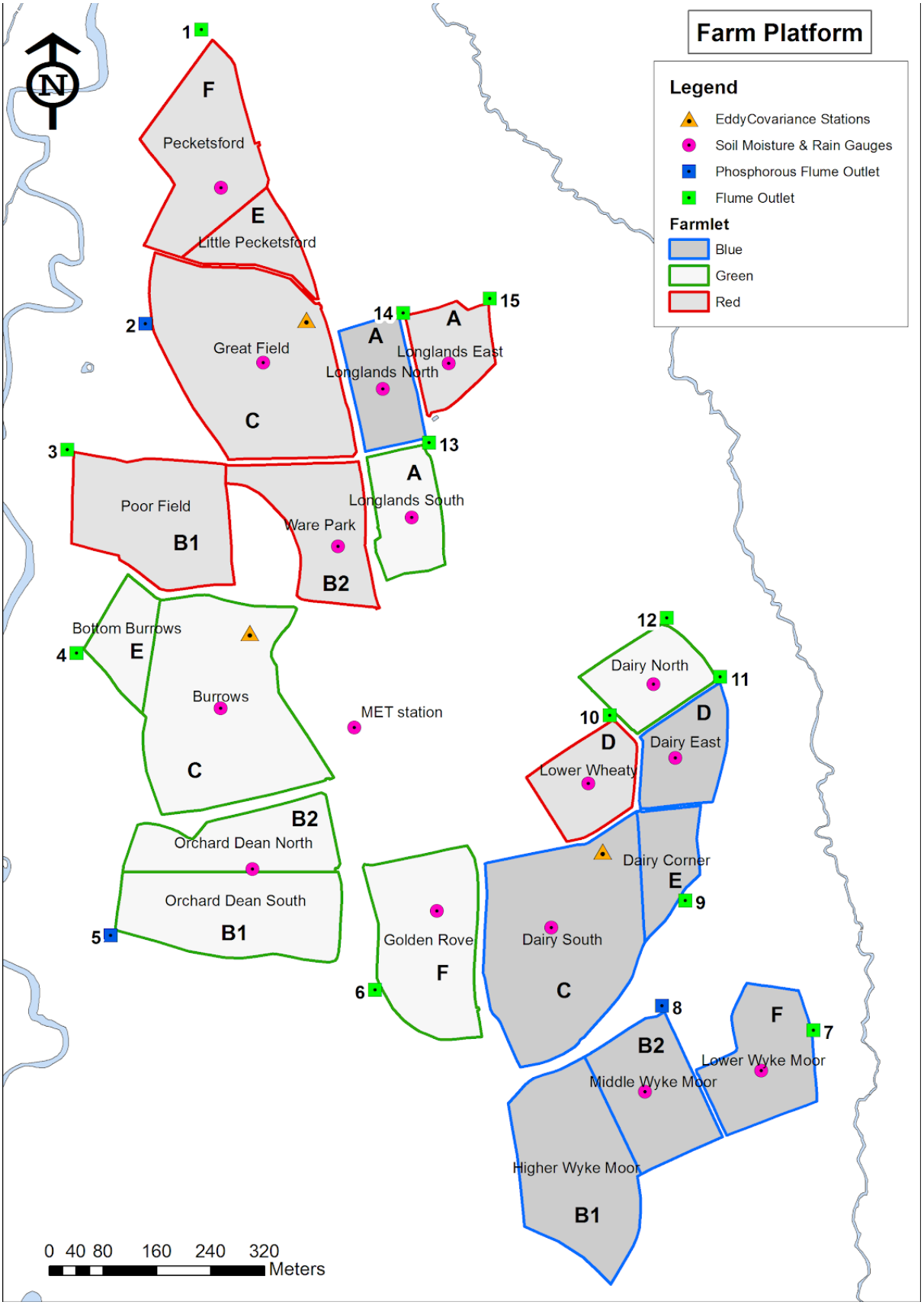


Figure 1-2: The North Wyke Farm Platform.

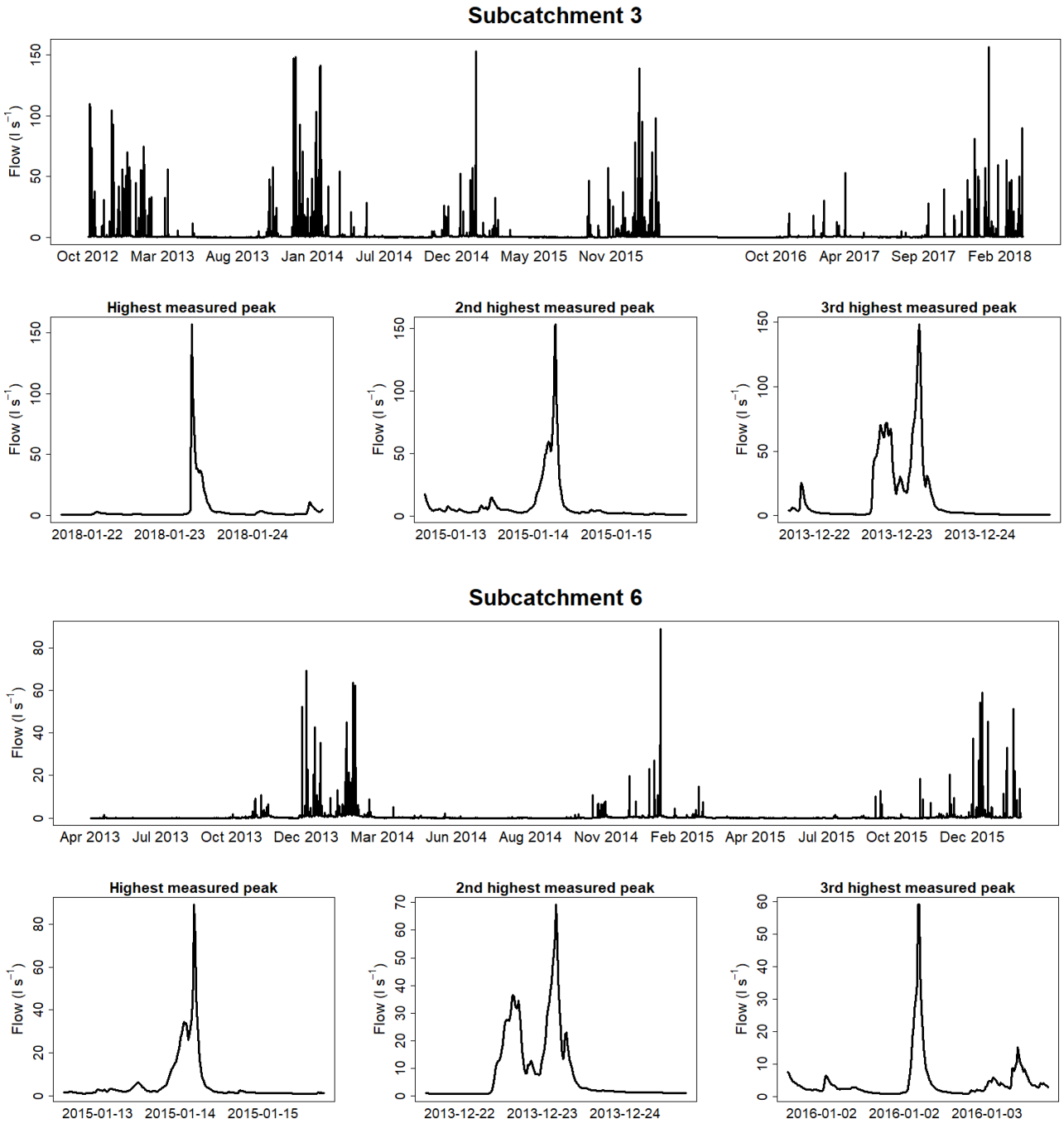


Figure 1-3: Measured 15-minutes flow at sub-catchments 3 and 6 along with the three highest peaks

1.8 Study process-based model

For this research, the 'SPACSYS' model was chosen to simulate the discharge for the two sub-catchments of the NWFP over the periods of interest. The SPACSYS model is a process-

based, field-scale model which simulates key agricultural processes such as plant growth and development, soil carbon (C) and nitrogen (N) cycling, water dynamics and heat transformation (Wu et al., 2007) (Figure 1-4). The main processes concerning plant growth are assimilation, respiration, water and N uptake, partitioning of photosynthate and N, N-fixation for legume plants and root growth. SPACSYS is a multidimensional model where the dimensions of most components have been reduced to decrease computation time and run the simulations within a reasonable time frame of minutes to hours and according to the capacity of common computers. The soil water, heat, C and N components are two-dimensional, combined with a three-dimensional root system sub-model.

There are three pools where water is stored, soil profile, soil surface and canopy. Rainfall partially falls to the surface and partially is temporarily stored in the canopy. If surface water cannot infiltrate to the soil profile, it creates runoff when exceeds a certain threshold. The water in the soil profile is depleted by plant uptake, soil evaporation, deep percolation and horizontal flows to field drains. The Richards' equation for water potential is used in SPACSYS to simulate water redistribution in a soil profile. The Brooks and Corey (1966) or van Genuchten (1980) models for the retention curve can be chosen in the model and in this study the latter was applied. The evapotranspiration was estimated by the Penman-Monteith equation. More details for soil water processes, such as flow in micropores, hysteresis and groundwater flow can be found in (Jansson, 1998). .

Site-specific input data for the simulations include weather variables from the North Wyke site, soil properties, field and grass management (e.g., fertiliser application dates and composition, reseeding, grazing and cutting dates), and initialization of the state variables (standing biomass and root distribution, soil water content and temperature distribution).

Previous simulations of water runoff, soil moisture and other agricultural processes for the sub-catchments of the NWFP using SPACSYS can be found in Liu et al. (2018), where a detailed explanation on the SPACSYS calibration is given. Simulated water fluxes and soil moisture were reported in Wu et al. (2016) and Liu et al. (2018). In both studies, the comparison between measured and simulated data showed that SPACSYS captures well the flow but commonly underestimates the magnitude of peak flow events during wet periods and overestimates the low flows and the soil moisture is overestimated, especially during the dry periods. A sensitivity analysis for 61 input parameters against 27 output variables showed that soil conditions and management had higher impact than the weather conditions (Shan et al., 2021).The thesis outputs are generic and not specific to the SPACSYS model and thesis outputs are similarly generic to catchment-based water processes other than those found on the NWFP.

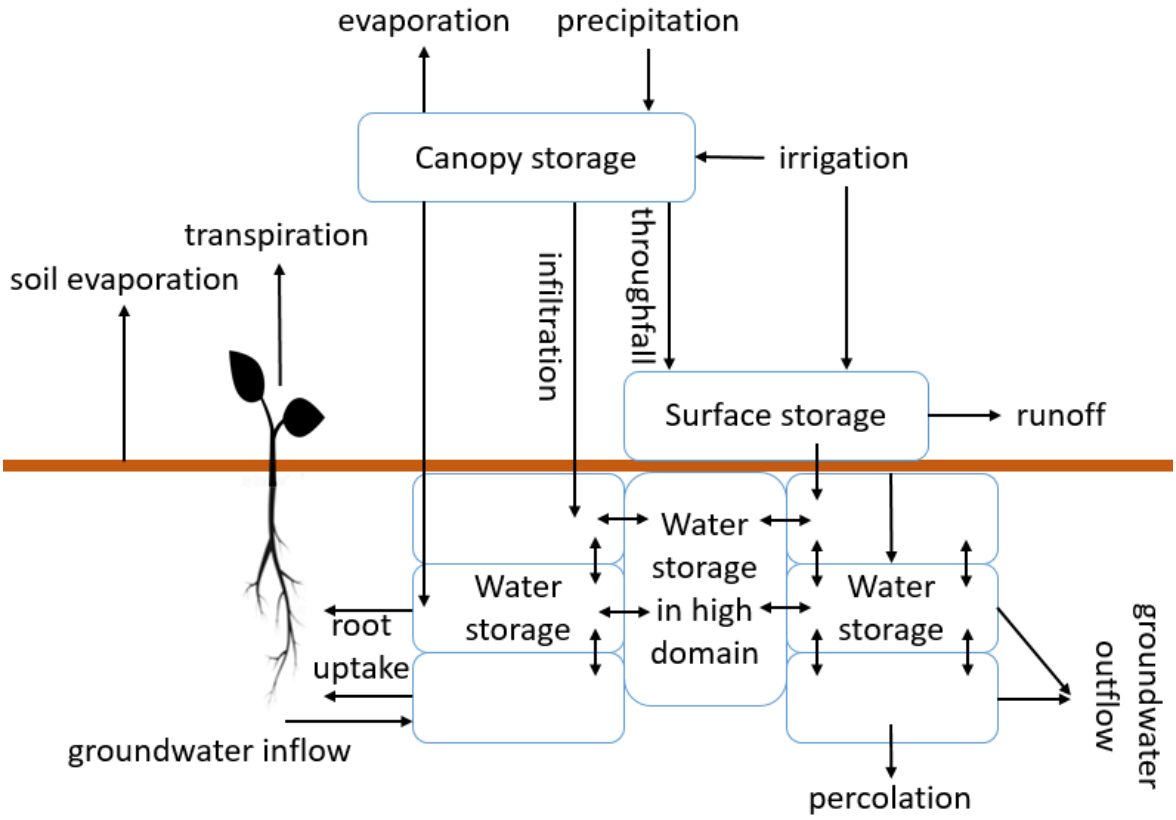


Figure 1-4: ：A conceptual diagram of water cycling in SPACSYS (Wu, et al., 2007; 2015). Water enters the soil as precipitation and then infiltrates into the soil where it flows down through the soil profile. The soil is conceptually made up as n layers in the profile (here three layers are shown).

1.9 Thesis aims, objectives and structure

The overall aim of this thesis was to explore techniques for improved modelling and forecasting of water flow, with a focus on an improved identification and characterisation of peak flow events across different temporal scales. Study outputs can be viewed as providing contribution to ongoing efforts in combining physical-based and statistical approaches into hybrid modelling but where the effects of scale are simultaneously considered. Four key research questions emerged:

Research question 1: How can a GPD be fitted to excesses above a threshold that is defined objectively?

Fitting the GP distribution to peak water flows involves several interconnected steps. First, the threshold over which peaks are considered extreme must be defined. This is important because it affects the sample size and the characteristics of the considered events, which in turn affects the parameter estimates and is impacted by the temporal scales. All this while the assumption of the independent and identically distributed observations must be respected. The dependence of the final fitted GDP on threshold selection means that it is important to have an objective and repeatable means of defining the threshold. In Chapter 2, I address this research question by developing a new automated process .

Research question 2: What is the effect of the time-step on the simulation of peak flow by a PBM?

To address this research question I explored the effect of temporal resolution and upscaling on the accuracy of peak flow obtained from the SPACSYS PBM. It provided new insights into the effects of simulating at various temporal scale on the identification of the frequency and magnitude of peak flow events.

Research question 3: Can statistically-based models of extremes be integrated with a PBM to improve the prediction of peak flows?

To answer this research question, the outputs from the PBM were post-processed using a statistical model that stems from Extreme Value Theory together with a machine learning model. This hybrid modelling framework (of 3 models) was assessed for its ability to improve the forecasting of peak discharge events and better capture the dynamics of the discharge process.

Research question 4: How much extra insight can the use of variograms and wavelets give us about models' performance?

All of the above research questions were answered by assessing the performance of the models through cross-validation of the point predictions together with associated accuracy indices. For this fourth study, the ability of variograms and wavelets to supplement this evaluation of models was investigated, with a focus on how well models could capture sources and changes in flow variance at different scales.

The four research questions are presented through Chapters 2 to 5, as a series of four published journal papers. Chapter 6 provides a detailed discussion of the research including future steps and concludes the research.

References

Abbott, M.B., J.C. Bathurst, J.A. Cunge, P.E. O'Connell and J. Rasmussen. (1986). An introduction to the European Hydrological System – Systeme Hydrologique Europeen, "SHE", 1: History and philosophy of a physically-based, distributed modelling system, Journal of Hydrology, 87(1-2). https://doi.org/45-59. 10.1016/0022-1694(86)90114-9.

Akhtar, M., N. Ahmad, and M. J. Booij, (2009). Use of regional climate model simulations as input for hydrological models for the Hindukush–Karakorum–Himalaya region, Hydrol. Earth Syst. Sci., 13, 1075-1089. https://doi.org/10.5194/hess-13-1075-2009.

Arnaud, P., J. Lavabre, C. Fouchier, S. Diss and P. Javelle. (2011) Sensitivity of hydrological models to uncertainty in rainfall input, Hydrological Sciences Journal, 56(3), 397-410. DOI: 10.1080/02626667.2011.563742

Arnold, J.G., R. Srinivasan, R.S. Muttiah and J.R. Williams. (1998). Large area hydrologic modeling and assessment part I: model development, J. Am. Water Resour. Assoc., 34(1), 73-89.

Asseng, S., F. Ewert, C. Rosenzweig, J. W. Jones, J. L. Hatfield, A. C. Ruane, K. J. Boote, et al. (2013). Uncertainty in Simulating Wheat Yields under Climate Change, Nature Climate Change, 3(9), 827. https://doi.org/10.1038/nclimate1916.

Bai, P., X. Liu, T. Yang, F. Li, K. Liang, S. Hu, and C. Liu. (2016). Assessment of the Influences of Different Potential Evapotranspiration Inputs on the Performance of Monthly Hydrological Models under Different Climatic Conditions, Journal of Hydrometeorology, 17(8), 2259-2274. https://doi.org/10.1175/JHM-D-15-0202.1.

Bárdossy, A. and S. K. Singh. (2008). Robust estimation of hydrological model parameters, Hydrol. Earth Syst. Sci., 12, 1273-1283. https://doi.org/10.5194/hess-12-1273-2008.

Bárdossy, A. and G. Pegram. (2017). Combination of Radar and Daily Precipitation Data to Estimate Meaningful Sub-Daily Point Precipitation Extremes, Journal of Hydrology, 544, 397-406. https://doi.org/10.1016/j.jhydrol.2016.11.039.

Bárdossy, András and G. G. S. Pegram. (2016). Space-Time Conditional Disaggregation of Precipitation at High Resolution via Simulation., Water Resources Research, 52(2), 920-937. https://doi.org/10.1002/2015WR018037.

Bates, B. C., Kundzewicz, Z. W., Wu, S. and Palutikof, J. P. (2008). Climate Change and Water. Technical Paper of the Intergovernmental Panel on Climate Change, IPCC Secretariat, Geneva, 210 pp.

Beechie, T. J., D. A. Sear, J. D. Olden, G. R. Pess, J. M. Buffington, H. Moir, P. Roni and M. M. Pollock. (2010). Process-Based Principles for Restoring River Ecosystems, BioScience, 60(3), 209-222. https://doi.org/10.1525/bio.2010.60.3.7.

Berliner, L. M. (2003). Physical-Statistical Modeling in Geophysics, Journal of Geophysical Research: Atmospheres, 108(D24), 8776. https://doi.org/10.1029/2002JD002865.

Beven, K. J. (1975). A deterministic, spatially distributed model of catchment hydrology. School of Environmental Sciences, University of East Anglia.

Beven, K. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. Adv. Water Resour, 16, 41-51. https://doi.org/10.1016/0309-1708(93)90028-E.

Bogner, K., K. Liechti and M. Zappa. (2016). Post-Processing of Stream Flows in Switzerland with an Emphasis on Low Flows and Floods, Water, 8(4), 115. https://doi.org/10.3390/w8040115.

Bogner, K., K. Liechti and M. Zappa. 2017. Technical Note: Combining Quantile Forecasts and Predictive Distributions of Streamflows, Hydrology and Earth System Sciences, 21(11), 5493-5502. https://doi.org/10.5194/hess-21-5493-2017.

Borra S. and A. Di Ciaccio. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Comput. Statist. Data Anal., 54 (12), 2976-2989. https://doi.org/10.1016/j.csda.2010.03.004.

Bouraoui, F., Grizzetti, B., Granlund, K., Rekolainen, S. and Bidoglio, G. (2004). Impact of Climate Change on the Water Cycle and Nutrient Losses in a Finnish Catchment, Climatic Change, 66(1-2), 109-126. doi: 10.1023/B:CLIM.0000043147.09365.e3.

Box, G. E. P. and G. M. Jenkins. (1976). Time Series Analysis. Forecasting and Control. Revised Edition. Holden-Day, San Francisco, CA.

Bradley, A. A., M. Habib and S. S. Schwartz. (2015). Climate Index Weighting of Ensemble Streamflow Forecasts Using a Simple Bayesian Approach, Water Resources Research, 51(9), 7382-7400. https://doi.org/10.1002/2014WR016811.

Breinl, K. and G. Di Baldassarre. (2019). Space-Time Disaggregation of Precipitation and Temperature across Different Climates and Spatial Scales, Journal of Hydrology: Regional Studies, 21, 126-146. https://doi.org/10.1016/j.ejrh.2018.12.002.

Brooks, R. H. and A. T. Corey. (1966). Properties of Porous Media Affecting Fluid Flow, Journal of the Irrigation and Drainage Division, 92(2).

Brown, I., R. Bardgett, P. Berry, I. Crute, J. Morison, M. Morecroft, J. Pinnegar, T. Reeder and K. Topp. (2016). UK Climate Change Risk Assessment - Chapter 3: Natural Environment and Natural Assets.

Brown, J. D. and D. J. Seo. (2010). A Nonparametric Postprocessor for Bias Correction of Hydrometeorological and Hydrologic Ensemble Forecasts, Journal of Hydrometeorology, 11(3), 642-665. https://doi.org/10.1175/2009JHM1188.1.

Caldeira, T. L., C. R. Mello, S. Beskow, L. C. Timm and M. R. Viola. (2019). LASH hydrological model: An analysis focused on spatial discretization, Catena, 173, 183–193.

Carpenter, T.M. and K.P. Georgakakos. (2006). Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales, J. Hydrol., 329(1), 174-185. https://doi.org/10.1016/j.jhydrol.2006.02.013.

Chen, H., C.-Y. Xu and S. Guo. (2012). Comparison and Evaluation of Multiple GCMs, Statistical Downscaling and Hydrological Models in the Study of Climate Change Impacts on Runoff, J. Hydrol. 434-435, 36-45. doi:10.1016/j.jhydrol.2012.02.040.

Chen, L., N. Sun, C. Zhou, J. Zhou, Y. Zhou, J. Zhang and Qing Zhou. (2018). Flood Forecasting Based on an Improved Extreme Learning Machine Model Combined with the Backtracking Search Optimization Algorithm, Water, 10(10), 1362. https://doi.org/10.3390/w10101362.

Choi, Y. S. , M. J. Shin and K. T. Kim. (2018). Preliminary Study of Computational Time Steps in a Physically Based Distributed Rainfall–Runoff Model, Water 10(9), 1269. https://doi.org/10.3390/w10091269.

Clarke, M. L. and H. M. Rendell. (2006). Hindcasting Extreme Events: The Occurrence and Expression of Damaging Floods and Landslides in Southern Italy, Land Degradation & Development, 17(4), 365-380. https://doi.org/10.1002/ldr.743.

Cloke, H. L. and F. Pappenberger. (2009). Ensemble Flood Forecasting: A Review, Journal of Hydrology, 375(3), 613-626. https://doi.org/10.1016/j.jhydrol.2009.06.005.

Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer, London, UK.

Collet, L., L. Beevers and C. Prudhomme. (2017). Assessing the Impact of Climate Change and Extreme Value Uncertainty to Extreme Flows across Great Britain, Water, 9(2), 103. https://doi.org/10.3390/w9020103.

Comber, A., C. Brunsdon, M. Charlton and P. Harris. (2017). Geographically Weighted Correspondence Matrices for Local Error Reporting and Change Analyses: Mapping the Spatial Distribution of Errors and Change, Remote Sensing Letters, 8(3), 234-243. https://doi.org/10.1080/2150704X.2016.1258126.

Crawford, N.H., Linsley, R.K., 1966. Digital simulation in hydrology, Stanford Watershed Model IV. Department of Civil Engineering, Stanford University, Technical Report 39.

Das, T, A. Bárdossy, E. Zehe and Y. He. (2008). Comparison of Conceptual Model Performance Using Different Representations of Spatial Variability, Journal of Hydrology, 356(1), 106-118. https://doi.org/10.1016/j.jhydrol.2008.04.008.

Do, H. X., S. Westra and M. Leonard. (2017). A Global-Scale Investigation of Trends in Annual Maximum Streamflow, Journal of Hydrology, 552, 28-43. https://doi.org/10.1016/j.jhydrol.2017.06.015.

Dogulu, N., López López, P., Solomatine, D. P., Weerts, A. H. and Shrestha, D. L. (2015). Estimation of Predictive Hydrologic Uncertainty Using the Quantile Regression and UNEEC Methods and Their Comparison on Contrasting Catchments, Hydrology and Earth System Sciences, 19 (7), 3181-3201. doi: 10.5194/hess-19-3181-2015.

Donelly-Makowecki, L.M. and R.D. Moore. (1999). Hierarchical testing of three rainfall-runoff models in small forested catchments, J. Hydrol., 219 (3–4), 136-152. https://doi.org/10.1016/j.ejrh.2019.100655.

El Hassan, A. A., H. O. Sharif, T. Jackson and S. Chintalapudi. (2013). Performance of a conceptual and physically based model in simulating the response of a semi-urbanized watershed in San Antonio, Texas, Hydrological Processes, 27(24), 3394-3408.

Elshorbagy, A., G. Corzo, S. Srinivasulu and D. P. Solomatine. 2010. Experimental Investigation of the Predictive Capabilities of Data Driven Modeling Techniques in Hydrology - Part 1: Concepts and Methodology, Hydrology and Earth System Sciences, 14(10), 1931-1941. https://doi.org/10.5194/hess-14-1931-2010.

Engeland, K., Hisdal, H. and Frigessi, A. (2004). Practical Extreme Value Modelling of Hydrological Floods and Droughts: A Case Study. Extremes, 7(1), 5–30.

Fathian, F., Mehdizadeh, S., Kozekalani A. S. and Safari, M. J. S. (2019). Hybrid Models to Improve the Monthly River Flow Prediction: Integrating Artificial Intelligence and Non-Linear Time Series Models, Journal of Hydrology, 575, 1200–1213. doi: 10.1016/j.jhydrol.2019.06.025.

Fatichi, S., E. R. Vivoni, F. L. Ogden, V. Y. Ivanov, B. Mirus, D. Gochis, C. W. Downer et al. (2016). An Overview of Current Applications, Challenges, and Future Trends in Distributed

Process-Based Models in Hydrology, Journal of Hydrology, 537, 45-60. https://doi.org/10.1016/j.jhydrol.2016.03.026.

Ficchì, A., Perrin, C. and Andréassian, V. (2016). Impact of temporal resolution of inputs on hydrological model performance: An analysis based on 2400 flood events, J. Hydrol., 538, 454-470. https://doi.org/10.1016/j.jhydrol.2016.04.016.

Field, C. B., Barros, V., Stocker, T. F. and Dahe, Q. (2012). Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change, Cambridge, Cambridge University Press.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample, Proc. Cambridge Philos. Soc., 24(2), 180-190.

Freeze, R. A. and R. L. Harlan. (1969). Blueprint for a Physically-Based, Digitally-Simulated Hydrologic Response Model, Journal of Hydrology, 9, 237-258. https://doi.org/10.1016/0022-1694(69)90020-1.

Guha-Sapir, D., P. Hoyois, P. Wallemacq and R. Below. (2016). Annual Disaster Statistical Review 2016: the numbers and trends.

Haddeland, I., D. P. Lettenmaier and T. Skaugen. (2006). Reconciling Simulated Moisture Fluxes Resulting from Alternate Hydrologic Model Time Steps and Energy Budget Closure Assumptions, Journal of Hydrometeorology, 7(3), 355-370. https://doi.org/10.1175/JHM496.1.

Haddeland, I., B. V. Matheussen and D. P. Lettenmaier. (2002). Influence of Spatial Resolution on Simulated Streamflow in a Macroscale Hydrologic Model, Water Resources Research, 38(7), 29-1-29-10. https://doi.org/10.1029/2001WR000854.

Hall, J., B. Arheimer, M. Borga, R. Brázdil, P. Claps, A. Kiss, T. R. Kjeldsen et al. (2014). Understanding Flood Regime Changes in Europe: A State-of-the-Art Assessment, Hydrology and Earth System Sciences, 18(7), 2735-2772. https://doi.org/10.5194/hess-18-2735-2014.

Hansen, J., M. Sato and R. Ruedy. (2012). Perception of Climate Change, Proceedings of the National Academy of Sciences, 109(37), E2415–23. https://doi.org/10.1073/pnas.1205276109.

Harris, P., C. Brunsdon and M. Charlton. (2013). The Comap as a Diagnostic Tool for Non-Stationary Kriging Models, International Journal of Geographical Information Science, 27(3), 511-541. https://doi.org/10.1080/13658816.2012.698014.

Harrod, T. R. and Hogan, D. V. (2008). The soils of North Wyke and Rowden. Soil Survey of England and Wales. pp. 1-54.

Hemri, S., D. Lisniak and B. Klein. (2015). Multivariate Postprocessing Techniques for Probabilistic Hydrological Forecasting, Water Resources Research, 51(9), 7436-7451. https://doi.org/10.1002/2014WR016473.

Hodgkins, G. A., P. H. Whitfield, D. H. Burn, J. Hannaford, B. Renard, K. Stahl, A. K. Fleig et al. (2017). Climate-Driven Variability in the Occurrence of Major Floods across North America and Europe, Journal of Hydrology, 552, 704-717. https://doi.org/10.1016/j.jhydrol.2017.07.027.

Huang, Y., A. Bárdossy and Ke Zhang. (2019). Sensitivity of Hydrological Models to Temporal and Spatial Resolutions of Rainfall Data, Hydrology and Earth System Sciences, 23(6), 2647-2663. https://doi.org/10.5194/hess-23-2647-2019.

Jaiswal, R., S. Ali and B. Bharti. (2020). Comparative evaluation of conceptual and physical rainfall–runoff models, J. Appl. Water Sci., 10, 1-14. https://doi.org/10.1002/hyp.9443

Jansson, P. E. (1998). Simulation model for soil water and heat conditions. Description of the SOIL model. Division of Agricultural Hydrotechnics Communications, 98:2, Swedish University of Agricultural Sciences, Uppsala.

Jenkinson, A. F. (1955). The Frequency Distribution of the Annual Maximum (or Minimum) Values of Meteorological Elements. Quarterly Journal of the Royal Meteorological Society, 81(348), 158-171.

Jeong, J., N. Kannan, J. Arnold, R. Glick, L. Gosselink and Raghavan Srinivasan. (2010). Development and Integration of Sub-Hourly Rainfall–Runoff Modeling Capability Within a Watershed Model, Water Resources Management, 24(15), 4505-4527. https://doi.org/10.1007/s11269-010-9670-4.

Jonkman, S. N. and J. K. Vrijling. (2008). Loss of Life Due to Floods, Journal of Flood Risk Management, 1(1), 43–56. https://doi.org/10.1111/j.1753-318X.2008.00006.x.

Kalman, R. E. 1960. A New Approach to Linear Filtering and Prediction Problems, Journal of Basic Engineering, 82(1), 35-45. https://doi.org/10.1115/1.3662552.

Kavetski, D., F. Fenicia and M. P. Clark. (2011). Impact of Temporal Data Resolution on Parameter Inference and Model Identification in Conceptual Hydrological Modeling: Insights from an Experimental Catchment, Water Resources Research, 47 (5). https://doi.org/10.1029/2010WR009525.

Khatami, S., C. Murray Peel, T. J. Peterson and A. W. Western. (2019). Equifinality and Flux Mapping: A New Approach to Model Evaluation and Process Representation Under

Uncertainty, Water Resources Research, 55(11), 8922-8941. https://doi.org/10.1029/2018WR023750

Kisi, O. and M. Cimen. 2011. A Wavelet-Support Vector Machine Conjunction Model for Monthly Streamflow Forecasting, Journal of Hydrology, 39(1), 132-140. https://doi.org/10.1016/j.jhydrol.2010.12.041.

Kobold, M. and K. Sušelj. (2005). Precipitation forecasts and their uncertainty as input into hydrological models, Hydrol. Earth Syst. Sci., 9, 322–332. https://doi.org/10.5194/hess-9-322-2005.

Kobold, M. and M. Brilly. (2006). The Use of HBV Model for Flash Flood Forecasting, Natural Hazards and Earth System Sciences, 6(3), 407-417. https://doi.org/10.5194/nhess-6-407-2006.

Kundzewicz, Z. W., Mata, L. J., Arnell, N. W., Doll, P., Kabat, P., Jimenez, B. et al. (2007). Freshwater Resources and Their Management. In Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by M. L. Parry, O. F. Canziani, J. P. Palutikof, P. J. van der Linden, and C. E. Hanson, 173–210. Cambridge University Press.

Lane, R. A., G. Coxon, J. E. Freer, T. Wagener, P. J. Johnes, J. P. Bloomfield, S. Greene, C. J. A. Macleod and S. M. Reaney. (2019). Benchmarking the Predictive Capability of Hydrological Models for River Flow and Flood Peak Predictions across over 1000 Catchments in Great Britain, Hydrology and Earth System Sciences, 23(10), 4011-4032. https://doi.org/10.5194/hess-23-4011-2019.

Langousis, A., A. Mamalakis, M. Puliga, and R. Deidda. (2016). Threshold Detection for the Generalized Pareto Distribution: Review of Representative Methods and Application to the NOAA NCDC Daily Rainfall Database. Water Resources Research, 52(4), 2659-2681. https://doi.org/10.1002/2015WR018502.

Li, W., Q. Duan, C. Miao, A. Ye, W. Gong, and Z. Di. (2017). A Review on Statistical Postprocessing Methods for Hydrometeorological Ensemble Forecasting. Wiley Interdisciplinary Reviews: Water, 4(6), e1246. https://doi.org/10.1002/wat2.1246.

Li, X.Q., J. Chen, C.Y. Xu, L. Li, and H. Chen. (2019). Performance of Post-Processed Methods in Hydrological Predictions Evaluated by Deterministic and Probabilistic Criteria. Water Resources Management, 33(9), 3289-3302. https://doi.org/10.1007/s11269-019-02302-y.

Lidén, R, and J Harlin. (2000). Analysis of Conceptual Rainfall–Runoff Modelling Performance in Different Climates. Journal of Hydrology, 238(3), 231-247. https://doi.org/10.1016/S0022-1694(00)00330-9.

Liu, Y., Y. Li, P. Harris, L. M. Cardenas, R. M. Dunn, H. Sint, P. J. Murray, M. R. F. Lee, and L. Wu. (2018). Modelling Field Scale Spatial Variation in Water Run-off, Soil Moisture, N2O Emissions and Herbage Biomass of a Grazed Pasture Using the SPACSYS Model. Geoderma, 315, 49-58. https://doi.org/10.1016/j.geoderma.2017.11.029.

López López, P., J. S. Verkade, A. H. Weerts, and D. P. Solomatine. (2014). Alternative Configurations of Quantile Regression for Estimating Predictive Uncertainty in Water Level Forecasts for the Upper Severn River: A Comparison. Hydrology and Earth System Sciences, 18(9), 3411-3428. https://doi.org/10.5194/hess-18-3411-2014.

McMillan, H. K., D. J. Booker, and C. Cattoën. (2016). Validation of a National Hydrological Model. Journal of Hydrology, 541, 800-815. https://doi.org/10.1016/j.jhydrol.2016.07.043.

Newman, A. J., M. P. Clark, K. Sampson, A. Wood, L. E. Hay, A. Bock, R. J. Viger, et al. (2015). Development of a Large-Sample Watershed-Scale Hydrometeorological Data Set for the Contiguous USA: Data Set Characteristics and Assessment of Regional Variability in Hydrologic Model Performance. Hydrology and Earth System Sciences, 19(1), 209-223. https://doi.org/10.5194/hess-19-209-2015.

Papacharalampous, G., H. Tyralis, A. Langousis, A. W. Jayawardena, B. Sivakumar, N. Mamassis, A. Montanari, and D. Koutsoyiannis. (2019). Probabilistic Hydrological Post-Processing at Scale: Why and How to Apply Machine-Learning Quantile Regression /Algorithms. Water, 11(10), 2126. https://doi.org/10.3390/w11102126.

Peck, S. L. (2004). Simulation as Experiment: A Philosophical Reassessment for Biological Modeling. Trends in Ecology & Evolution, 19(10), 530-534. https://doi.org/10.1016/j.tree.2004.07.019.

Pickands, J. (1975). Statistical Inference Using Extreme Order Statistics. The Annals of Statistics, 3(1), 119-131. http://www.jstor.org/stable/2958083.

Pui, A., A. Sharma, R. Mehrotra, B. Sivakumar, and E. Jeremiah. (2012). A Comparison of Alternatives for Daily to Sub-Daily Rainfall Disaggregation. Journal of Hydrology, 470-471, 138-157. https://doi.org/10.1016/j.jhydrol.2012.08.041.

Quilty, J., J. Adamowski, and M.A. Boucher. (2019). A Stochastic Data-Driven Ensemble Forecasting Framework for Water Resources: A Case Study Using Ensemble Members

Derived From a Database of Deterministic Wavelet-Based Models. Water Resources Research, 55(1), 175-202. https://doi.org/10.1029/2018WR023205.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Monthly Weather Review, 133(5), 1155-1174. https://doi.org/10.1175/MWR2906.1.

Roulin, E., and S. Vannitsem. (2011). Postprocessing of Ensemble Precipitation Predictions with Extended Logistic Regression Based on Hindcasts. Monthly Weather Review, 140(3), 874-888. https://doi.org/10.1175/MWR-D-11-00062.1.

Rust, W., R. Corstanje, I. P. Holman, and A. E. Milne. (2014). Detecting Land Use and Land Management Influences on Catchment Hydrology by Modelling and Wavelets. Journal of Hydrology, 517, 378-389. https://doi.org/10.1016/j.jhydrol.2014.05.052.

Scarrott, C., and A. MacDonald. (2012). A Review of Extreme Value Threshold Es-Timation and Uncertainty Quantification. REVSTAT–Statistical Journal, 10(1), 33-60.

Schaake, J., K. Franz, A. Bradley, and R. Buizza. (2006). The Hydrologic Ensemble Prediction EXperiment (HEPEX). Hydrology and Earth System Sciences Discussions, 3(5), 3321-3332. https://hal.archives-ouvertes.fr/hal-00298784.

Schaake, J. C., V. I. Koren, Q.Y. Duan, K. Mitchell, and F. Chen. (1996). Simple Water Balance Model for Estimating Runoff at Different Spatial and Temporal Scales. Journal of Geophysical Research: Atmospheres, 101(D3), 7461-7475. https://doi.org/10.1029/95JD02892.

Schumann, A. H. (1993) Development of conceptual semi-distributed hydrological models and estimation of their parameters with the aid of GIS, Hydrological Sciences Journal, 38:6, 519-528, DOI: 10.1080/02626669309492702.

Shan, Y., M. Huang, P. Harris and L. Wu. (2021). A Sensitivity Analysis of the SPACSYS Model, Agriculture, 11(7), 624. https://doi.org/10.3390/agriculture11070624

Sikorska, A. E., A. Montanari, and D. Koutsoyiannis. (2015). Estimating the Uncertainty of Hydrological Predictions through Data-Driven Resampling Techniques. Journal of Hydrologic Engineering, 20(1), A4014009. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000926.

Smith, P., J. U. Smith, D. S. Powlson, W. B. McGill, J. R. M. Arah, O. G. Chertov, K. Coleman, et al. (1997). A Comparison of the Performance of Nine Soil Organic Matter Models Using Datasets from Seven Long-Term Experiments. Geoderma, Evaluation and Comparison of Soil Organic Matter Models, 81(1, 153-225. https://doi.org/10.1016/S0016-7061(97)00087-6.

Solomatine, D. P., and A. Ostfeld. (2008). Data-Driven Modelling: Some Past Experiences and New Approaches. Journal of Hydroinformatics, 10(1), 3-22. https://doi.org/10.2166/hydro.2008.015.

Takahashi, T., P. Harris, M. S. A. Blackwell, L. M. Cardenas, A. L. Collins, J. a. J. Dungait, J. M. B. Hawkins, et al. (2018). Roles of Instrumented Farm-Scale Trials in Trade-off Assessments of Pasture-Based Ruminant Production Systems. Animal, 12(8), 1766-1776. https://doi.org/10.1017/S1751731118000502.

Thibault, K. M., and J. H. Brown. (2008). Impact of an Extreme Climatic Event on Community Assembly. Proceedings of the National Academy of Sciences, 105(9), 3410-3415. https://doi.org/10.1073/pnas.0712282105.

Toth, E., A. Montanari, and A. Brath. (1999). Real-Time Flood Forecasting via Combined Use of Conceptual and Stochastic Models. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere, 24(7), 793-798. https://doi.org/10.1016/S1464-1909(99)00082-9.

Tsutsumida, N., P. Rodríguez-Veiga, P. Harris, H. Balzter, and A. Comber. (2019). Investigating Spatial Error Structures in Continuous Raster Data. International Journal of Applied Earth Observation and Geoinformation, 74, 259-268. https://doi.org/10.1016/j.jag.2018.09.020.

Uusitalo, L., A. Lehikoinen, I. Helle, and K. Myrberg. (2015). An Overview of Methods to Evaluate Uncertainty of Deterministic Models in Decision Support. Environmental Modelling & Software, 63, 24-31. https://doi.org/10.1016/j.envsoft.2014.09.017.

van Genuchten, M. Th. (1980). A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils, Soil Science Society of America Journal, 44(5), 892-898. https://doi.org/10.2136/sssaj1980.03615995004400050002x

Wasko, C., A. Sharma, and D. P. Lettenmaier. (2019). Increases in Temperature Do Not Translate to Increased Flooding. Nature Communications, 10(1), 5676. https://doi.org/10.1038/s41467-019-13612-5.

Wijayarathne, D. B., and P. Coulibaly. (2020). Identification of Hydrological Models for Operational Flood Forecasting in St. John's, Newfoundland, Canada. Journal of Hydrology: Regional Studies, 27, 100646. https://doi.org/10.1016/j.ejrh.2019.100646.

Wu, L., M. B. McGechan, N. McRoberts, J. A. Baddeley, and C. A. Watson. (2007). SPACSYS: Integration of a 3D Root Architecture Component to Carbon, Nitrogen and Water Cycling—Model Description. Ecological Modelling, 200(3), 343-359. https://doi.org/10.1016/j.ecolmodel.2006.08.010.Wu, L., A. P. Whitmore and G. Bellocchi. (2015). Modelling the impact of environmental changes on grassland systems with SPACSYS. Advances in Animal Biosciences, 6, 37-39. doi:10.1017/S2040470014000508.

Yaseen, Z. M., O. Jaafar, R. C. Deo, O. Kisi, J. Adamowski, J. Quilty, and A. El-Shafie. (2016). Stream-Flow Forecasting Using Extreme Learning Machines: A Case Study in a Semi-Arid Region in Iraq. Journal of Hydrology, 542, 603-614. https://doi.org/10.1016/j.jhydrol.2016.09.035.

Yen, H., X. Wang, D.G. Fontane, R.D. Harmel and M. Arabi. (204). A framework for propagation of uncertainty contributed by parameterization, input data, model structure, and calibration/validation data in watershed modelling, Environ. Model. Softw., 54, 211-221. https://doi.org/10.1016/j.envsoft.2014.01.004.

Yu B, Ciesiolka C. A. A, Rose C. W and Coughlan K. J. (1997). Note on sampling errors in the rainfall and runoff data collected using tipping bucket technology. Transactions of the American Society of Agricultural Engineers 40, 1305-1309.

Zhao, L., Q. Duan, J. Schaake, A. Ye, and J. Xia. (2011). A Hydrologic Post-Processor for Ensemble Streamflow Predictions. In Advances in Geosciences, 29,51-59. Copernicus GmbH. https://doi.org/10.5194/adgeo-29-51-2011.

Zhou, J., T. Peng, C. Zhang, and N. Sun. (2018). Data Pre-Analysis and Ensemble of Various Artificial Neural Networks for Monthly Streamflow Forecasting. Water, 10(5), 628. https://doi.org/10.3390/w10050628.

2. An evaluation of automated GPD threshold selection methods for hydrological extremes across different scales

Stelian Curceac[a,*], Peter M. Atkinson[b, c, d, e], Alice Milne[f], Lianhai Wu[a], Paul Harris[a]

[a] Rothamsted Research, Department of Sustainable Agriculture Sciences, North Wyke EX20 2SB, Devon, UK.

[b] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK.

[c] Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

[d] School of Geography, Archaeology and Palaeoecology, Queen's University Belfast, BT7 1NN, Northern Ireland, UK.

[e] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China.

[f] Rothamsted Research, Department of Sustainable Agriculture Sciences, Harpenden AL5 2JQ, UK

[*]Corresponding author: stelian.curceac@rothamsted.ac.uk

Published in Journal of Hydrology

2.1 Abstract

This study investigated core components of an extreme value methodology for the estimation of high-flow frequencies from agricultural surface water run-off. The Generalized Pareto distribution (GPD) was used to model excesses in time-series data that resulted from the 'Peaks Over Threshold' (POT) method. First, the performance of eight different GPD parameter estimators was evaluated through a Monte Carlo experiment. Second, building on the estimator comparison, two existing automated GPD threshold selection methods were evaluated against a proposed approach that automates the threshold stability plots. For this second experiment, methods were applied to discharge measured at a highly-instrumented agricultural research facility in the UK. By averaging fine-resolution 15-minute data to hourly, 6-hourly and daily scales, we were also able to determine the effect of scale on threshold selection, as well as the performance of each method. The results demonstrate the advantages of the proposed threshold selection method over two commonly applied methods, while at the same time providing useful insights into the effect of the choice of the scale of measurement on threshold selection. The results can be generalized to similar water monitoring schemes and are important for improved characterizations of flood events and the design of associated disaster management protocols.

Keywords: Generalized Pareto Distribution; Peaks over threshold; Threshold selection; Flood Frequency Analysis; Scale effects; Grassland agriculture.

2.2 Introduction

The magnitude and frequency of floods is likely to increase as a result of climate change (Bates et al., 2008; Field et al., 2012; Kundzewicz et al., 2007) and this could push ecosystems beyond the threshold of normal disturbance resulting in negative impacts that may be irreversible (e.g. Thibault & Brown, 2008). Increased surface run-off intensify erosion and introduce more soil, organic matter and pollutants into water courses. Floods in areas of steep and unstable slopes increase the possibility of landslides (Clarke & Rendell, 2006). Moreover, increased water runoff generally results in higher sediments and nutrient losses that can lead to soil degradation (Bouraoui et al., 2004). Flooding can have severe impacts on key ecosystem services, such as those of support (e.g. water, nutrient cycling and soil protection), regulation (e.g. climate) and culture (e.g. scenic recreation) (MA, 2005).

Flood Frequency Analysis (FFA) is a classic method to analyze the relationship between flood magnitude and the corresponding frequency of occurrence. Reliable estimation and prediction of high flow quantiles require extrapolation beyond the observed range of events, commonly using parametric probability distributions. There are two main approaches for defining extreme events in stationary time-series. The first is the block (usually annual) maxima (AM) method where the dataset is divided into contiguous blocks of equal size and the maximum values in each segment are considered. According to the Fisher-Tippet theorem (Fisher & Tippett, 1928), these identically, independently distributed (iid) random variables asymptotically follow a Generalized Extreme Value (GEV) distribution (Coles, 2001; Jenkinson, 1955). The second approach is known as the peaks-over threshold (POT) method, which considers the values X that exceed a fixed high threshold u. The distribution function of the excess values $X - u$, conditional on $X > u$, is a Generalized Pareto Distribution (GPD)

(Pickands, 1975). The case study we consider, contains six years of fine resolution (15-minute) flow measurements, which is insufficient for effective fitting of the GEV distribution. Therefore, only the POT method with the GPD was investigated.

The above two families of distributions have fundamental differences, but also theoretical links (see Langousis et al., 2016). The GEV distribution is usually best fitted to annual maxima samples and for this reason long historic records are required. This restriction does not apply to the POT method since it includes all the peaks above a certain threshold allowing for greater flexibility. The threshold must be large enough for the excesses to follow a GPD, but an over-estimated threshold leads to reduced sample size and increases the variance of the estimates. A smaller threshold increases the sample size but also the bias of the estimates as the empirical distribution deviates from a perfect GPD model (Scarrott and MacDonald, 2012). Clearly, GPD threshold selection is of key importance and there is no universally recognized best performing method although various techniques have been proposed (see e.g. Langousis et al. 2016 and Scarrott & MacDonald, 2012). Among them are probabilistic-based techniques (Beirlant et al., 1996, 2006; Choulakian & Stephens, 2001; Deidda & Puliga, 2006; Goegebeur et al., 2008; Hill, 1975), computational approaches (Beirlant et al., 2005; Danielsson et al. 2001; Hall, 1990; Thompson et al., 2009; Zoglat et al., 2014) and mixture models (Behrens et al., 2004; Eastoe & Tawn, 2010; Solari & Losada, 2012). Graphical methods (Das & Ghosh, 2013; Deidda, 2010; Lang et al., 1999; Tanaka & Takara, 2010), such as the Mean Residual Life (MRL) plot (Coles 2001; Beguería, 2005; Davison & Smith, 1990) are used commonly for the selection of an optimal threshold, but have been criticized for the difficulty and subjectivity of their interpretation (Scarrott & MacDonald 2012; Yang et al., 2018). Alternatively, analytical methods have the advantage that they can be automated, and the associated

uncertainty can be quantified. Solari et al. (2017) proposed an automated threshold selection method based on AD goodness of fit test. The application of their technique on long records of precipitation and flow resulted in estimated thresholds that were within the stability regions of the shape and modified scale parameters. Durocher et al. (2018) compared several automatic methods and proposed a hybrid one where consistency with shape stability was found for most of the considered sites.

In this study, we propose an empirical automated method for threshold determination, based on threshold stability, which is evaluated against two commonly applied analytical methods, together with eight alternatives for GPD parameter estimation. Furthermore, by averaging the case study's 15-minute flow data to hourly, 6-hourly and daily supports, we determine the effects of temporal measurement scale on threshold selection, as well as the performance of each method.

The remainder of this paper is organized as follows. Section 2.3 presents the methods for GPD parameter estimation, two analytical threshold selection techniques, this study's proposed automated threshold stability method, and model evaluation diagnostics and indices. Section 2.4 describes the case study site and flow data, together with the simulation experiment design used to evaluate the performance of the different GPD parameter estimators. Results are presented in Section 2.5, which includes an investigation of scale effects through a series of flow data integrations. Sections 2.6 and 6 discuss and conclude the study, respectively.

2.3 Methodology

The cumulative distribution function (CDF) of the iid excesses over an appropriate threshold u for the GPD is:

$$G(x) = \Pr(X - u < x | X > u) = \begin{cases} 1 - \left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-\frac{1}{\xi}}, \xi \neq 0 \\ 1 - e^{\left(-\frac{x-u}{\sigma}\right)}, \quad \xi = 0 \end{cases} \qquad 2\text{-}1$$

where x, for this study, is the extreme flow in m^3s^{-1}, u is the location parameter, σ is the scale parameter and ξ is the shape parameter. The value of the shape parameter defines the type of distribution from the GPD family, that is, $\xi = 0$ refers to the exponential distribution, for $\xi > 0$ the corresponding distribution has a heavy upper tail that behaves like a power function with exponent $-1/\xi$ and for $\xi = 1$ the distribution is uniform. The Pareto distribution is obtained when $\xi < 0$ (Figure 2-1).

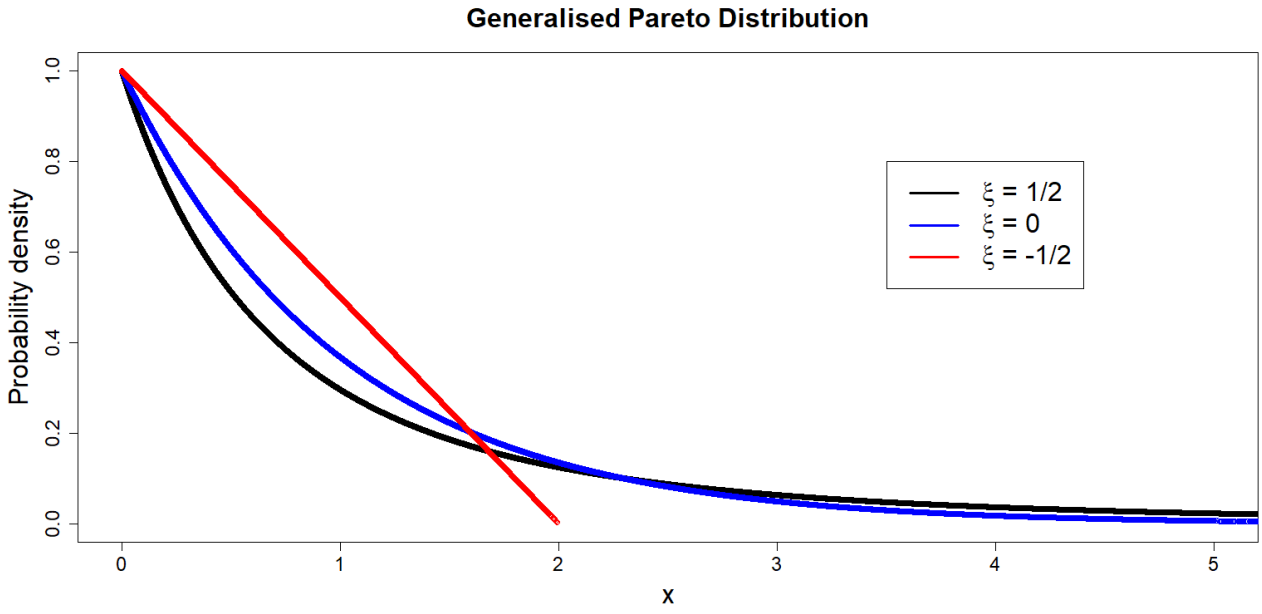


Figure 2-1: GPD for different values of the shape parameter ξ.

2.3.1 GPD parameter estimators

The excesses above a suitable threshold are modelled by the GPD and the parameters of the distribution can be estimated by competing methods, where the Maximum Likelihood estimator (MLE) is the most commonly used (Prescott & Walden, 1980, 1983; Smith, 1985).

Hosking and Wallis (1987) showed that MLE provides greater variance and bias for small samples compared to the Probability Weighted Moment (PWM) (Greenwood et al., 1979; Landwehr et al., 1979) and the Method of Moments (MOM) estimators. Coles and Dixon (1999) proposed a modified MLE which contains a penalty function for the shape parameter (i.e. the Maximum Penalized Likelihood estimator (MPLE). Zhang (2007) presented a hybrid Likelihood Moment estimator (LME) which provides feasible estimates and has high asymptotic efficiency. All of these methods are evaluated in this study, together with that suggested by Pickands (1975) and a maximum goodness-of-fit (MGF) estimator (e.g. Luceño, 2006). Estimator performance has been found to depend significantly on sample size and the value of the GPD shape parameter (Ashkar & Tatsambon, 2007; de Zea Bermudez & Kotz, 2010; Hosking & Wallis, 1987), and the choice of the estimator should be made based on the specifics of the situation. The equations for the above estimators can be found in Appendix A.

2.3.2 Threshold selection methods

The selection of the threshold u is a crucial step in GPD extreme value analysis. On the one hand, a small threshold results in a large sample that makes statistical inference more effective, but can lead to biased estimates due to deviations of the empirical distributions from the GPD model (e.g. Beirlant et al., 2005). On the other hand, when considering large thresholds and consequently small samples, parameter estimates have a smaller expected bias, but a larger variance that can be highly dependent on the estimation method. The two main approaches for threshold selection are graphical methods, such as the MRL plot, and analytical methods that can be automated.

An important assumption for the application of the POT method is that the extracted peaks are independent. A commonly applied method is to use no more than 2-3 peaks per year (Madsen et al., 1997; Todorovic, 1978) but it has been criticised for lack of flexibility. Another solution is to consider a minimum separation interval between successive peaks (Cunnane, 1979; Lang et al., 1999). This minimum separation interval accords to the scale and nature of the measured process, but for daily flow data, an interval of a few days commonly ensures that the peaks are generated from different events (Engeland et al., 2004). The autocorrelation function is a popular choice for the investigation of serial dependence in a time series. However, this approach assumes normally distributed variables, which is not the case for peak discharges, so other independence tests should be implemented (e.g. Ledford and Tawn, 2003; Reiss and Thomas, 2007). In this study, and through prior experimentation, maximum peaks separated by a minimum of three days were considered and their independence was tested using Kendall's τ test (Claps and Laio, 2003; Ferguson et al., 2000).

2.3.2.1 Graphical methods: MRL plots

The most popular graphical method is the MRL plot (Coles, 2001; Davison & Smith, 1990). If the scaled excesses $X_{u^*} = [X - u^* | X > u^*]$ above a threshold u^* are Generalized Pareto (GP) distributed, then for every $u \geq u^*$, the scaled excesses $X_u = [X - u | X > u]$ are similarly GP distributed with the same shape parameter ξ, a scale parameter $\sigma_u = \sigma_{u^*} + \xi(u - u^*)$ and a mean value:

$$\overline{X}(u) = E[X - u | X > u] = \frac{\sigma_u}{1 - \xi} = \frac{\sigma_{u^*} + \xi(u - u^*)}{1 - \xi} = Au + B \qquad 2\text{-}2$$

where $A = \xi/(1 - \xi)$ and $B = (\sigma_{u^*} - \xi u^*)/(1 - \xi)$ are the respective slope and intercept of the linear relation. The sample estimates of the mean excesses are then plotted for

different values of the threshold and the most appropriate is considered to be the one after which the mean excesses follow a straight line (e.g. Das & Ghosh, 2013).

Another graphical technique is to plot the estimated shape and/or modified scale parameters for different threshold candidates and select the one above which the estimates are constant (Brodin & Rootzén, 2009; Bommier, 2014; Sigauke & Bere, 2017). The main criticism of graphical methods is that the interpretation of the plot can be ambiguous or subjective as it is usually unclear which part of the curve is linear (Scarrott & MacDonald, 2012). In this respect, attempts have been made to automate (Langousis et al., 2016) and estimate the uncertainty (Liang et al., 2019) of the graphical methods.

2.3.2.2 Analytical methods: Square Error and Normality of Differences

The Square Error (SE) method was developed by Zoglat et al. (2014) following the work of Beirlant et al. (2005), and is implemented as follows. Let $u_1, u_2, \dots, u_n$ be n equally spaced increasing threshold candidates. For each of these thresholds, estimate the scale σ_{u_j} and shape ξ_{u_j} parameters for $j = 1, \dots, n$. Find N_{u_j} the exceedances that correspond to each threshold u_j and simulate m independent samples of size N_{u_j} from the GPD with parameters σ_{u_j} and ξ_{u_j}. For each probability $a \in A = \{0.05, 0.1, \dots, 0.95\}$ and each $i = 1, \dots, m$ calculate the quantiles q^i_{a,u_j} and compute $q^{sim}_{a,u_j} = \frac{1}{m}\sum_{i=1}^{m} q^i_{a,u_j}$. The optimal threshold is the one for which the square error $SE_{u_j} = \sum_{a \in A}\left(q^{sim}_{a,u_j}/q^{obs}_{a,u_j}\right)^2$ between the simulated and the observed quantiles is minimum. The selection of the threshold candidates u_j can be defined by the user or as an automated process. For example, the smallest threshold can be set as zero or the median and the maximum threshold set as a high percentile of the data.

An alternative analytical method for threshold selection was proposed by Thompson et al. (2009). Again, let $u_1, u_2, \dots, u_n$ be n equally spaced increasing threshold candidates. For the excesses above the threshold u_j, $\hat{\sigma}_{u_j}$ and $\hat{\xi}_{u_j}$ are the MLEs of the scale and shape parameters, respectively, for $j = 1, \dots, n$. If $u \leq u_{j-1} < u_j$ is an appropriate threshold then according to Coles (2001), $\sigma_{u_{j-1}} = \sigma_u + \xi(u_{j-1} - u)$ and $\sigma_{u_j} = \sigma_u + \xi(u_j - u)$. Consequently, $\sigma_{u_j} - \sigma_{u_{j-1}} = \xi(u_j - u_{j-1})$ and from standard maximum likelihood theory we have that $E[\hat{\sigma}_{u_j}] \approx \sigma_{u_j}$ and $E\left[\hat{\xi}_{u_j}\right] = \xi$ for any j such that $u_j > u$. Respectively, $E\left[\tau_{u_j} - \tau_{u_{j-1}}\right] \approx 0$, $j = 2, \dots, n$ for $\tau_{u_j} = \hat{\sigma}_{u_j} - \hat{\xi}_{u_j} u_j$, $j = 1, \dots, n$. It follows that $\tau_{u_j} - \tau_{u_{j-1}}$ approximately follows a normal distribution. Thompson et al. (2009) suggest Pearson's Chi-square test to examine the null hypothesis of normality. However, this test has been criticised for having inferior power properties (Moore, 1986). For this reason, we also applied the Anderson-Darling, Cramer-von Mises, Kolmogorov-Smirnov and Shapiro-Francia normality tests (Thode, 2002). Regardless of which of the five normality tests are used, we refer to this method as the 'Normality of Differences' method. According to this approach, a suitable threshold $u \leq u_{j-1} < u_j$ is the one for which all the differences $\tau_{u_j} - \tau_{u_{j-1}}$ are approximately normally distributed. We selected the appropriate threshold as the one for which the *p*-value of $\tau_{u_j} - \tau_{u_{j-1}}, j = 2, \dots, n$ is above 0.05. A smaller threshold would be selected for a smaller *p*-value (e.g. 0.01).

2.3.2.3 Proposed method based on Threshold Stability

For this study, we propose an automated threshold selection method based on stability plots (Coles, 2001; Scarrott & MacDonald 2012). If the GPD is an appropriate model for the excesses above a threshold u, then for all larger thresholds $u^* > u$ it will also be suitable with the

shape parameter being relatively constant. In other words, it is the approximately linear horizontal part on the shape parameters versus thresholds plot. This does not apply for the scale parameter σ_{u^*}, as it changes with the threshold $\sigma_{u^*} = \sigma_u + \xi(u^* - u)$. However, the modified scale parameter $\sigma_1 = \sigma_{u^*} - \xi u$ remains relatively constant. Therefore, we fit a cubic smoothing spline to this plot and calculate the rate of change at each of m consecutive steps. The cubic smoothing spline estimate $\hat{f}$ of a function f in the model $Y_i = f(x_i) + \varepsilon_i$, is defined as the minimizer of $\sum_{i=1}^{n}\{Y_i - \hat{f}(x_i)\}^2 + \lambda \int \hat{f}''(x)^2 dx$, where λ is the smoothing parameter. The minimum change rate locates the part of the plot where the shape and the modified scale parameters reach a plateau.

A preliminary analysis showed that a smoothing parameter value of λ = 0.4 of the cubic spline function was the most appropriate to avoid both over- and under-fitting. A total of $n = 1000$ threshold candidates were used in each case and a cubic spline was fitted to the corresponding estimated shape and modified scale parameters. The numbers of the consecutive steps for which the minimum change rate was calculated, were $m = 25, 50, 75$ and 100 which corresponds to 2.5%, 5%, 7.5% and 10%, respectively, of the total number of fitted values, that is, the total threshold candidates n.

2.3.3 Evaluation procedure

Quantile-Quantile (Q-Q) plots are commonly used to investigate the efficiency of the statistical inference of the fitted GPD models. To quantify the difference between the theoretical and empirical quantiles for probabilities $a \in A = \{0.95, 0.951, \ldots, 0.999\}$, various error and agreement diagnostics were calculated. Specifically, we calculated the Mean Square Error (MSE) (e.g. Turan and Yurdusev, 2009), the Normalized Root Mean Square Error (NRMSE) (e.g. Sheta and El-Sherif, 1999) and the Relative Index of Agreement ($RD \in [0,1]$)

(Krause et al., 2005; Willmott, 1981). For ideal model performance, both MSE and NRMSE should tend to zero, while RD should tend to unity. The NRMSE was obtained by dividing the root MSE by the difference between minimum and maximum values and, thus, was less sensitive to very large values and provided a more robust diagnostic than MSE.

2.4 Study site and datasets

2.4.1 Study site

Discharge data come from a single sub-catchment of the North Wyke Farm Platform (NWFP). The NWFP is a farm-scale experiment established in 2011 in the southwest of England (50°46'10"N, 3°54'05"W) for research into sustainable grassland livestock systems (Orr et al., 2016; Takahashi et al., 2018). The platform is located at an altitude in the range of 120-180 m above sea level. The platform's fields have a declining slope at the west towards the River Taw and to the east, to one of its tributaries, the Cocktree stream. The soil texture consists of a slightly stony clay loam topsoil (approximately 36% clay) above a mottled stony clay (approximately 60% clay). The subsoil is impermeable to water and during rain events most of the excess water moves by surface and sub-surface lateral flow towards the drainage system described below.

Each of the 15 NWFP sub-catchments are hydrologically isolated through a combination of topography and a network of French drains (800 mm deep trenches), which ensure that the total runoff is channeled to instrumented flumes, measuring 15-minute water discharge and water chemistry from October 2012. The discharge from each sub-catchment is measured through a combination of primary and secondary flow devices (Liu et al., 2018). The primary devices are H-type flumes (TRACOM Inc., Georgia, USA) with capacity designed for a 1-in-50 year storm event. The specific design of the H-type flume facilitates the accurate

measurement of both low and high flows and is relatively self-cleaning since it allows the ready passage of sediment and particulate matter. A secondary flow measurement device (OTT hydromet, Loveland, CO., USA) is used to measure the stage within the flume and convert it to discharge rate using flume-specific formulae which depend on water height. The flow is generated only from rainfall as the fields are not irrigated. In each sub-catchment, 15-minute precipitation and soil moisture are also monitored. Detailed information on the NWFP are given in Section 1.7.

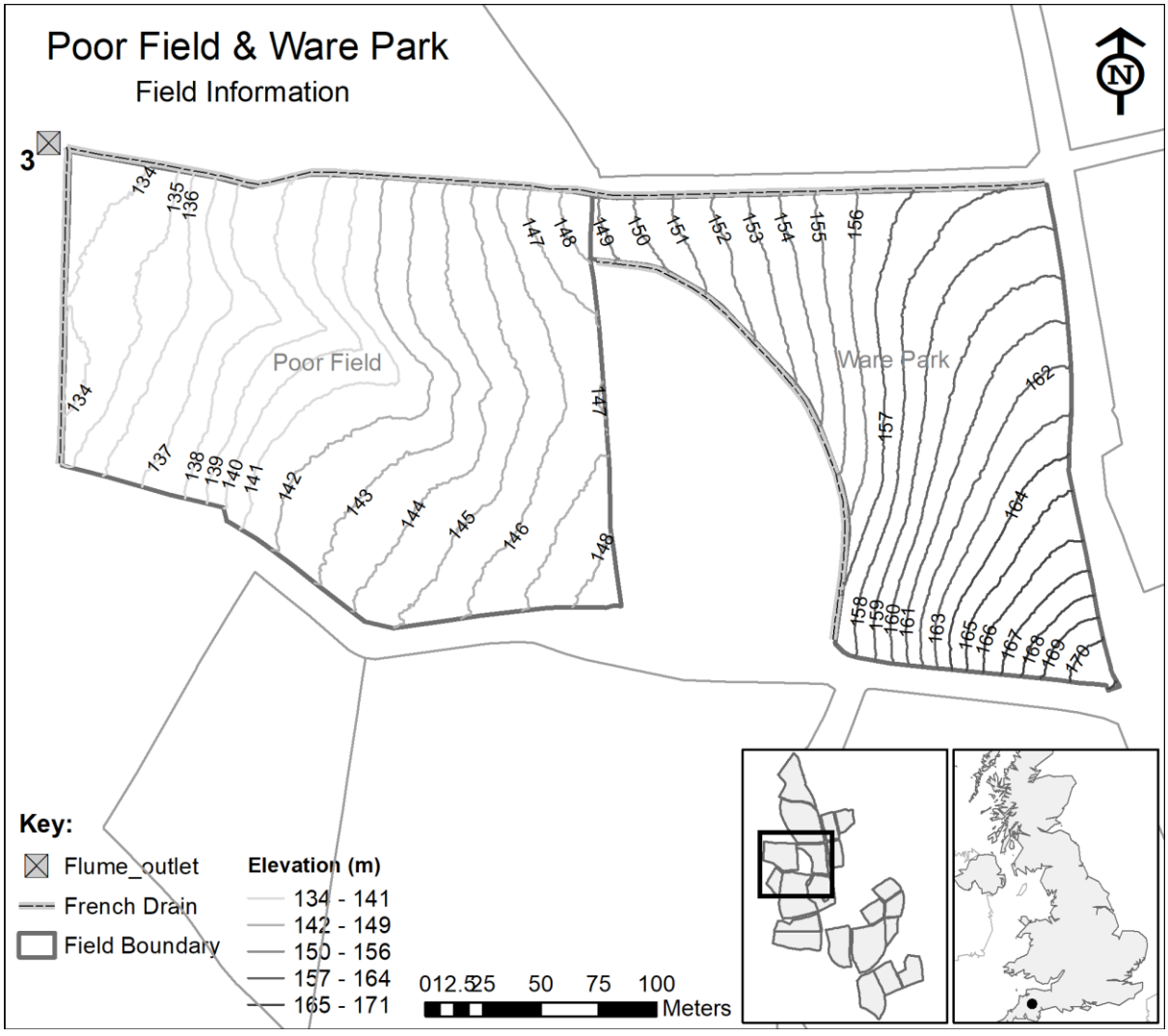


Figure 2-2: Details of the NWFP sub-catchment selected for this study (sub-catchment number 3 of 15, consisting of two fields called Poor Field and Ware Park). The Rain gauge is approximately centrally located in Ware Park (see Section 1.7, Figure 1-2).

2.4.2 Measured data

For this study, we used the flow discharge measured at sub-catchment 3 of the NWFP, which is part of the 'red' farmlet (Figure 2-2) and 6.84 ha in size. Given this is a methodological-based study, we chose to use data from this sub-catchment as it has one of the smallest number of missing values (approximately 1%) for the six-year period (2012-2018). Imputation of the missing values was performed using a regularized iterative Principal Components Analysis (PCA impute) model (Josse & Husson, 2013). The largest imputed value was approximately 20 l s^{-1} which is smaller than any threshold suggested (see below) and, therefore, is not considered as a peak flow and does not affect the subsequent analysis. It should be noted that, compared with measurements from many river or stream monitoring systems, the flow data (Figure 2-3) are highly discontinuous with many zeros, as non-zero measurements occur only after rainfall events.

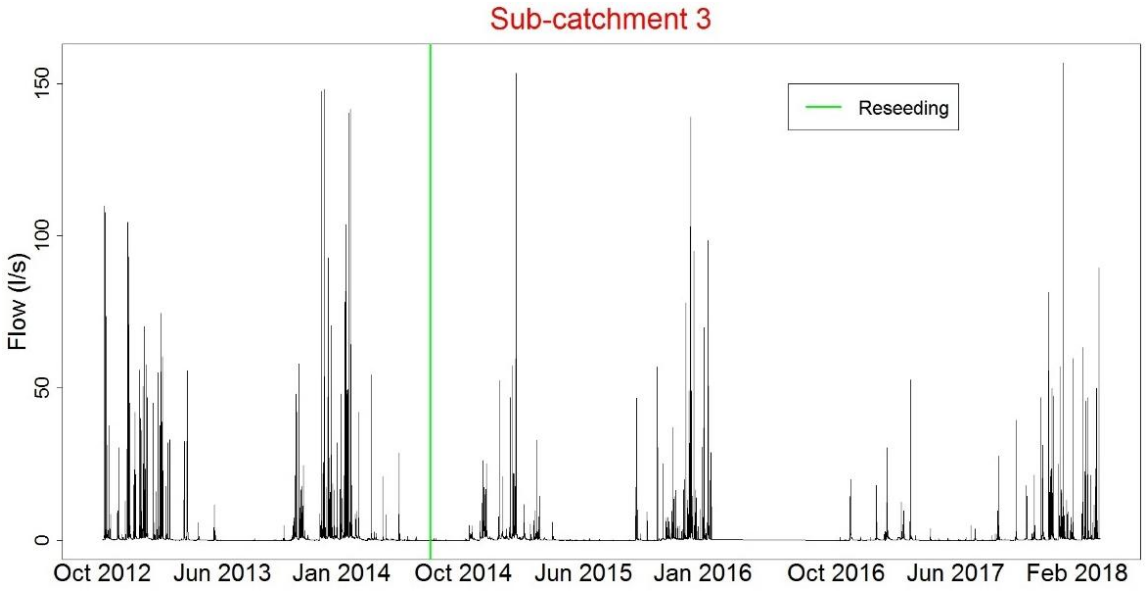


Figure 2-3: Flow (l s^{-1}) measurements at sub-catchment 3 (2012 to 2018).

2.4.3 Simulated data

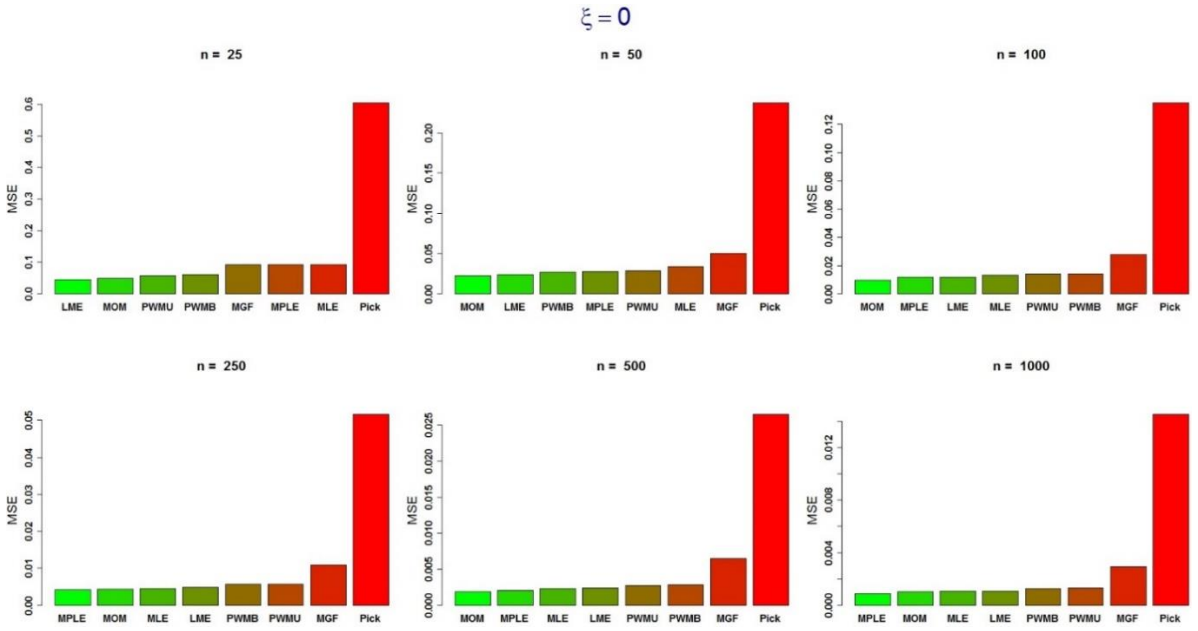
As a precursor to the empirical study, the performance of the eight GPD parameter estimators was assessed through a Monte Carlo experiment. We generated random time-series of different sample sizes ($n = 25,\ 50, 100, 250, 500, 1000$) from a GPD distribution with a known shape parameter (ξ = -0.5, -0.25, 0, 0.25 and 0.5). For each combination, 10,000 random samples were generated. The performance of the estimators was evaluated using: (a) bar plots for MSE values and (b) boxplots for estimated ξ. Here the "error" in MSE is the difference between the actual (or known) ξ and that estimated, where MSE incorporates both the variance and the bias of the estimators. Outcomes were used to guide the analyses with the measured NWFP flow data.

2.5 Results

2.5.1 Monte Carlo study for Performance of GPD estimators

Our simulated data analysis showed that the performance of the GPD parameter estimators depends on both the sample size n (see performance plots in Figure 2-4 for a shape parameter of $\xi = 0$ only) and the value of the shape parameter ξ (see Appendix B for performance plots with ξ = -0.5, -0.25, 0.25 and 0.5), which accords with previous studies (e.g. Gharib et al., 2017; Mackay et al., 2011). On viewing all plots, the maximum likelihood (MLE and MPLE) estimators were both negatively biased for small sample sizes for any value of the shape parameter and their performance increased in terms of bias and variance as sample size increased. The MLE outperformed the other estimators for large sample sizes for all values of the shape parameter. The unbiased and biased probability weighted moments, PWMU and PWMB respectively, were consistently the least biased amongst all estimators and provided a small variance, which was less sensitive to sample size compared to the

likelihood estimators. According to the MSE, the PWM estimators were most appropriate for small sample sizes and positive shape parameters. The MOM estimator had a similar behavior to the PWMs when $\xi \leq 0$ but had a negative bias for $\xi > 0$ and the bias increased as the value of the shape parameter and the sample size increased. Pickland's estimator ('Pick') and the MGF estimators produced a large variance and the least accurate estimates of the shape parameter, through the whole range of the examined values. LME was among the best performing estimators regarding accuracy and bias, except for the very short tails ($\xi = 0.5$, see Appendix B), when the estimates deviated greatly from the rest of the estimators and the predefined value of the shape parameter. In summary, the MLE/MPLE, PWMU/PWMB and the LME were considered the most unbiased and precise estimators and so we select only from this reduced group of estimators in subsequent analyses using the measured data.



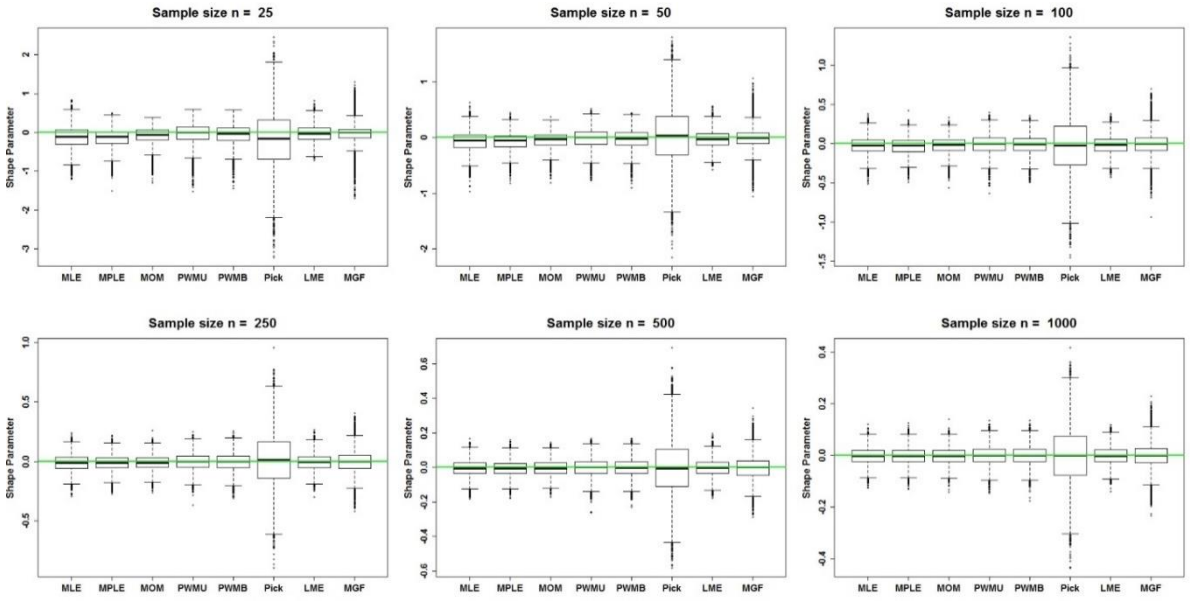


Figure 2-4: Comparison of the performance of GPD estimators for shape parameter ξ = 0 and for six different sample sizes ($n = 25,\ 50, 100, 250, 500, 1000$) evaluated by a Monte Carlo experiment.

2.5.2 Empirical study for Threshold Selection

2.5.2.1 Preliminary effects of data aggregation

Initially, the flow (l s^{-1}) time-series of 15-minute resolution was averaged to time-series data of 30 minutes, hourly, 3-hourly, 6-hourly, 12-hourly and daily resolutions. Figure 2-5 shows the behavior of the MLE-estimated shape parameters for a range of thresholds for the differently aggregated flow data. The range of thresholds was set from the median to the maximum for which daily flow can be fitted efficiently. The shape parameter is in the range of 0.5 to almost 2 for the minimum threshold, has a decreasing trend as the threshold increases and can become negative for the largest thresholds. The similar shape characteristics could be an indication that the shape parameter describes an inherent feature of the process and that changes of scale, which affect the size or variability of the observed values of the process, do not substantially change the shape characteristics of these observations. For the remainder of this study, results from the 30-minute, 3-hourly and 12-

hourly aggregations are not reported as retained aggregations (hourly, 6-hourly and daily) communicate all key outcomes adequately.

Kendall's τ test showed that the maximum peaks separated by a minimum of three days were reasonably independent (Figure 2-6). The statistics τ are large for the lowest thresholds where the peaks are numerous and autocorrelated. With an increasing threshold, the values of the τ decrease rapidly and are below the 95% acceptance limits which supports the null hypothesis of independence of the peaks.

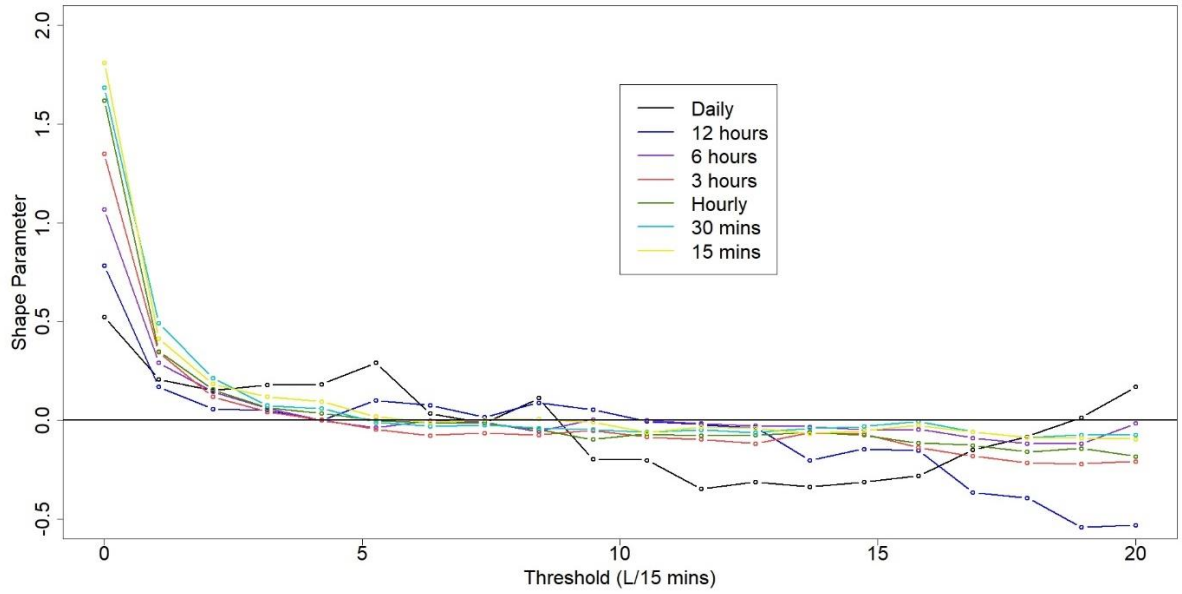


Figure 2-5: Shape parameter characteristics of measured (15-minute) and a series of averaged (30-minute to daily) flow rates.

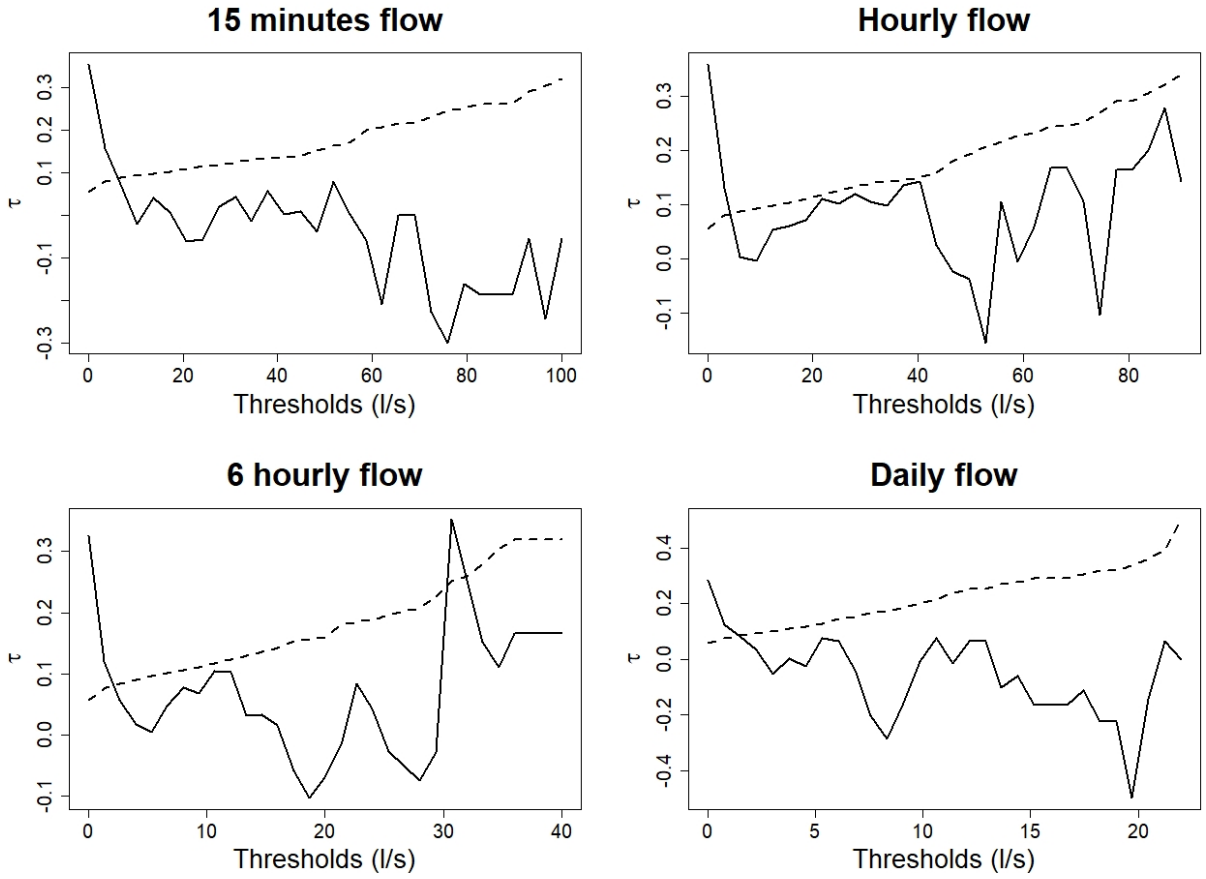


Figure 2-6: Kendall's test statistic τ (solid lines) along with the 95% acceptance limits of the test (dashed lines) of the independence of the maximum peaks separated by a minimum of three days.

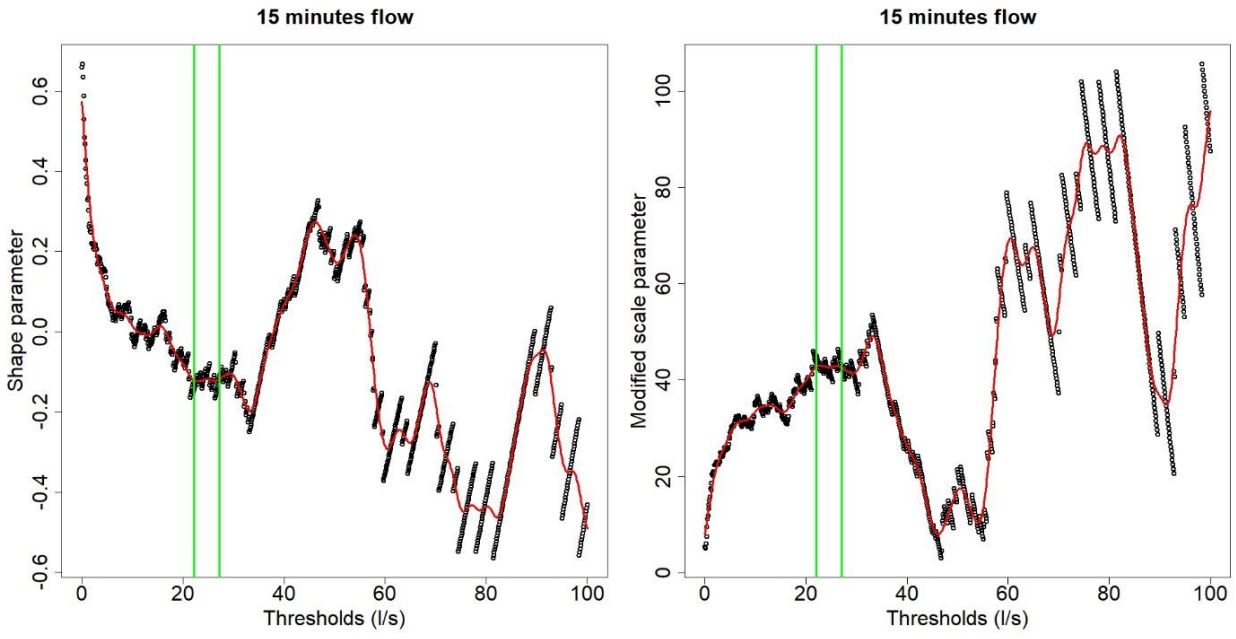
2.5.2.2 Automated Threshold Stability plots

The choice of estimators for the shape and modified scale parameters was guided by the results of the Monte Carlo experiment (Section 2.5.1). For example, for thresholds $u_j = 1,2,\ldots,5$ of the 15-minute flow data, the number of exceedances was $N_{u_j} > 300$ and the shape parameter ξ_{u_j} between 0.5 and 0.25. For this combination, MLE, MPLE, PWMU, PWMB and LME were the best performing estimators. Thus, for our empirical study, we choose LME due to its consistently precise and unbiased estimates of positive shape parameters for a large sample size. Increasing the thresholds u_j resulted in a reduced sample size ($100 < N_{u_j} < 250$) and negative values of the shape parameter. In this case, we choose MPLE for our

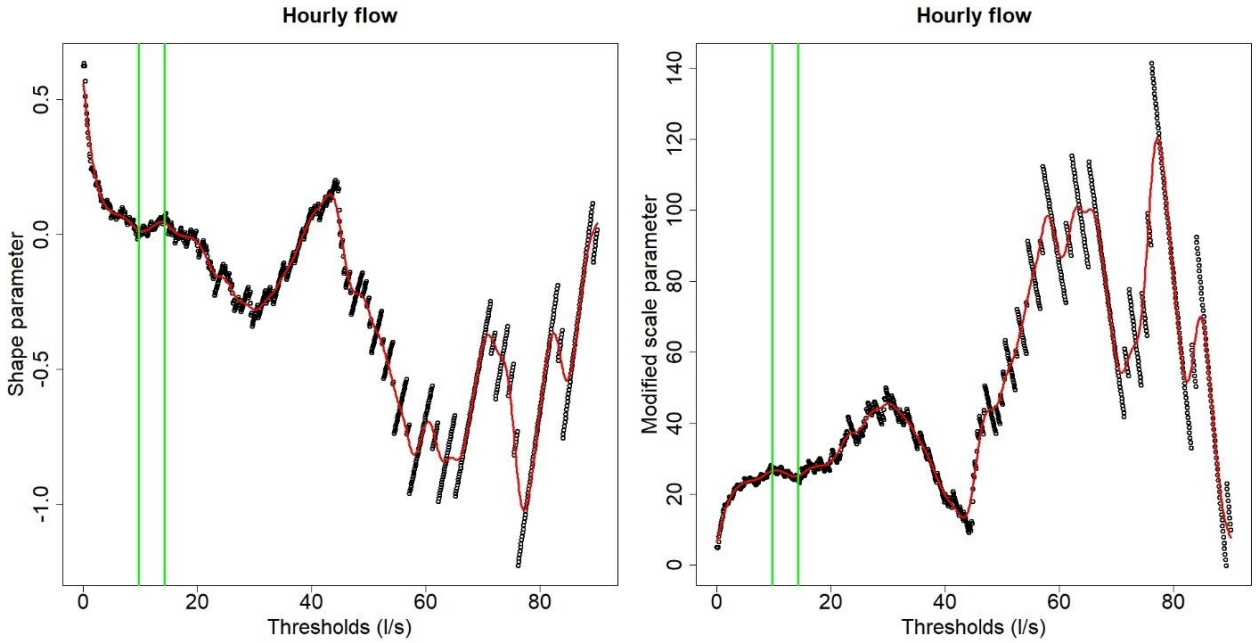
empirical work. In all the other cases, the PWMU estimator was preferred as it provided unbiased estimates with small variance.

Stability plots are given in Figure 2-7 for different flow aggregations, where results reveal our 'Automated Threshold Stability' (ATS) extension to be reasonably robust, since changes in the number of consecutive steps m had a very small impact on the selected threshold and usually resulted in over-lapping regions from which the threshold was considered. The peak flows at 15 minutes and hourly resolution did not provide many regions that could be considered as a plateau, so the number of consecutive steps was set to m = 50 (5% of the total) to also capture the smaller approximately linear horizontal parts. Interestingly, for each aggregation, fitting the same cubic spline functions to both the estimated shape and modified scale parameters, resulted in almost identical suggested thresholds.

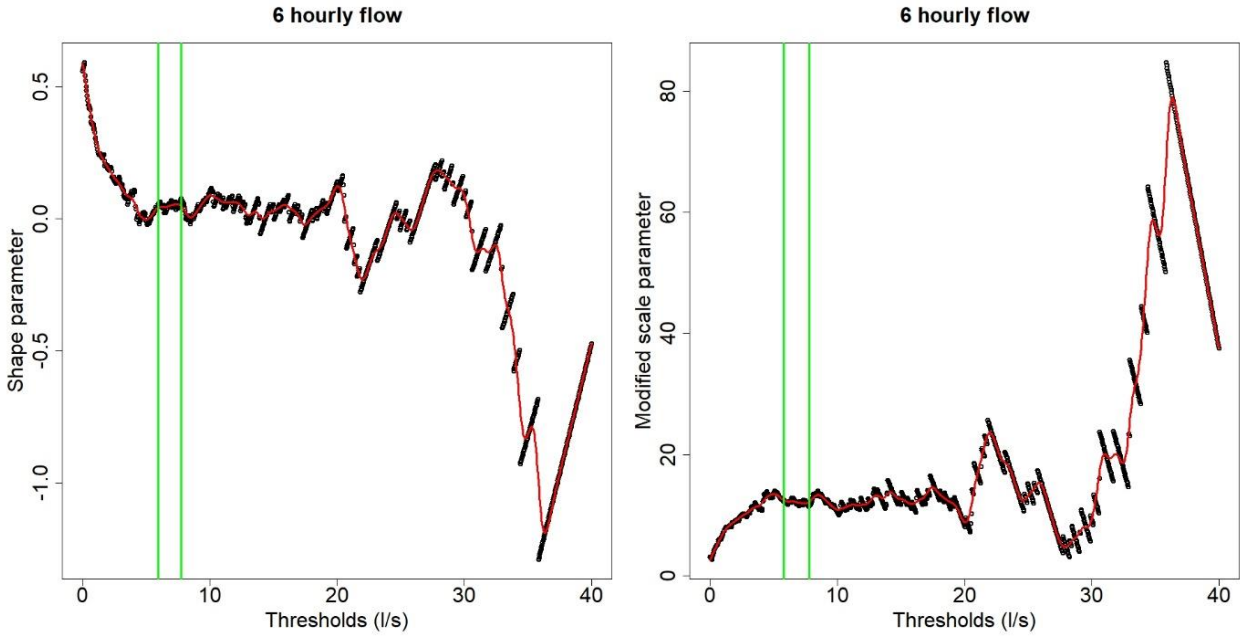
a)



b)


Hourly flow
Shape parameter
0.5
0.0
-0.5
-1.0
0
20
40
60
80
Thresholds (l/s)
Hourly flow
Modified scale parameter
0
20
40
60
80
100
120
140
0
20
40
60
80
Thresholds (l/s)

c)


6 hourly flow
Shape parameter
0.5
0.0
-0.5
-1.0
0
10
20
30
40
Thresholds (l/s)
6 hourly flow
Modified scale parameter
0
20
40
60
80
0
10
20
30
40
Thresholds (l/s)

d)

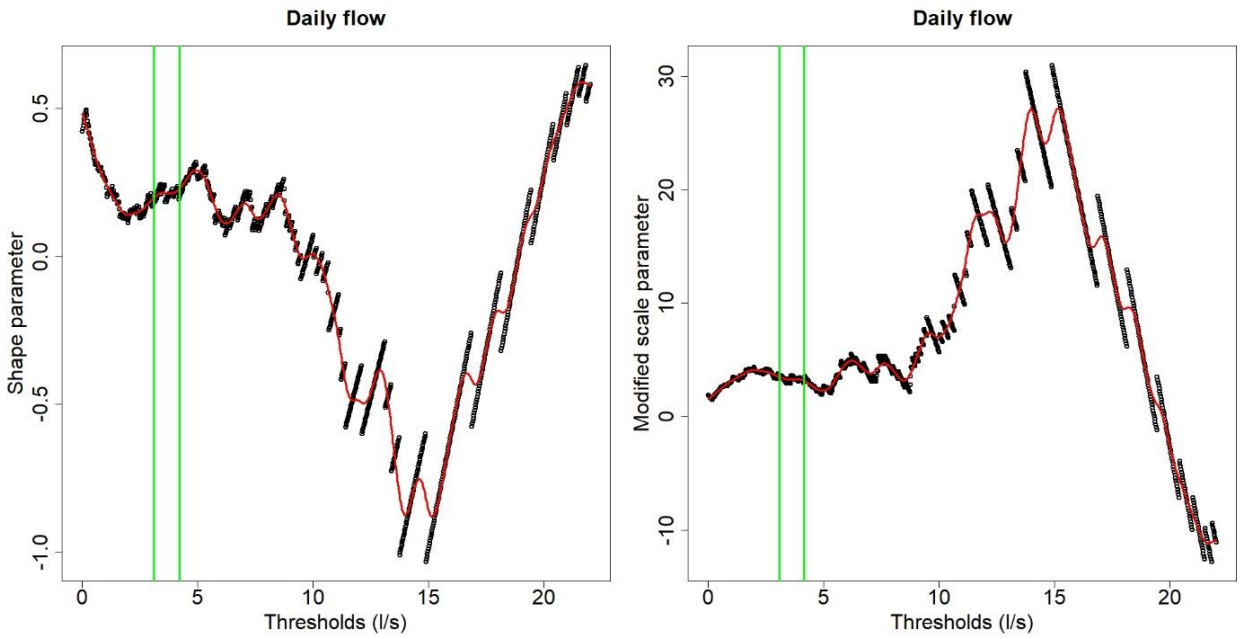


Figure 2-7: Automated Threshold Stability (ATS) method: Selected threshold (that between the vertical green lines) of a) 15 minutes, b) hourly, c) 6 hourly and d) daily flow based on smoothing splines.

2.5.2.3 Analytical threshold selection methods: Square Error and Normality of Differences

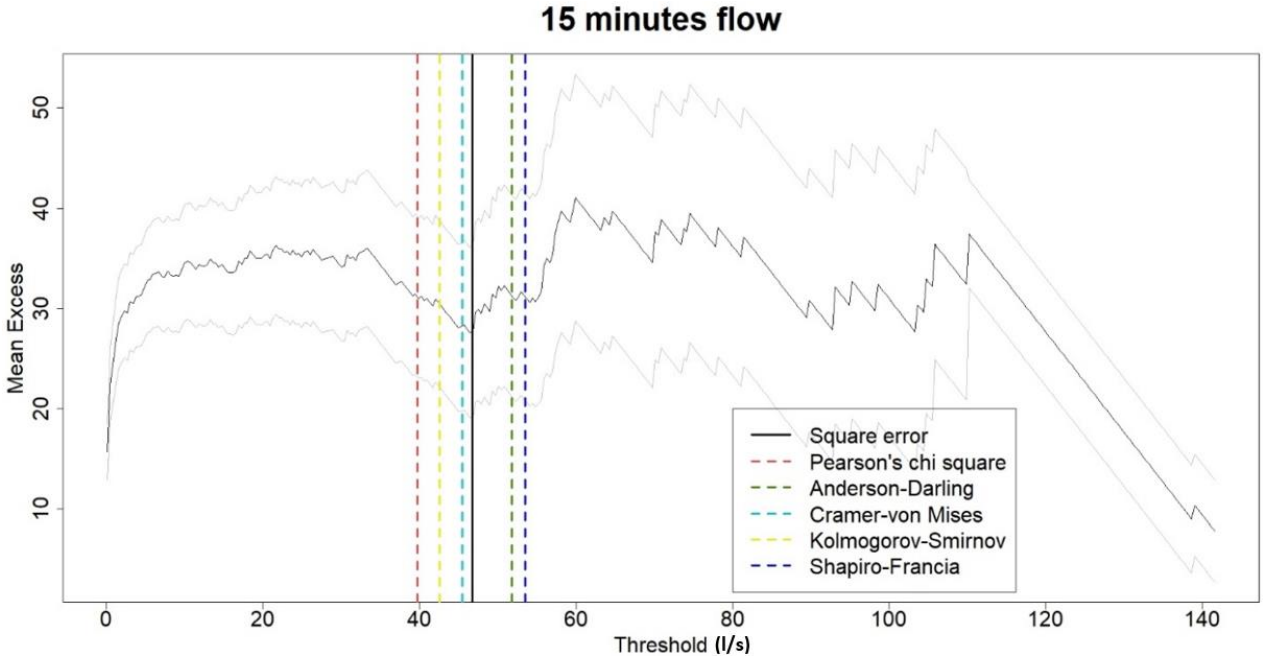
The choice of GPD estimators for the simulation of the quantiles for the SE method was performed using a similar procedure as described in Section 2.5.2.2, while the approach based on the Normality of Differences test is based on assumptions of maximum likelihood theory, and consequently the shape parameter was estimated by the MLE. The number n of the considered thresholds u_n plays an important role in the results. Thompson et al. (2009) suggested $n = 100$ and reported that for $n < 100$, less reliable results were obtained. We similarly specified $n = 100$ but also found the thresholds to be over-estimated for $n > 100$.

Our results indicated little consistency in the selection of thresholds where a specific part of the MRL plot could be considered approximately linear. The thresholds of the 15-minute peak

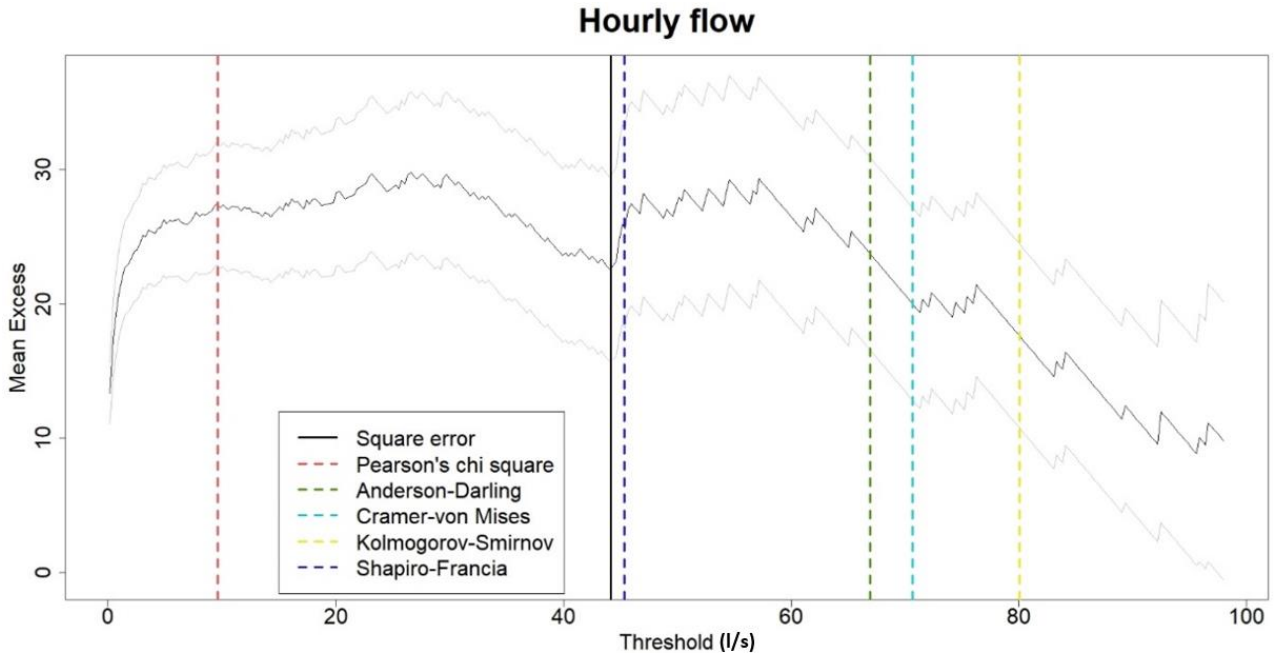
flow estimated by the SE method and the Normality of Differences tests (Figure 2-8a) are considerably larger than that based on this study's ATS method (Figure 2-7a) at around 40 to 50 l/s and 20 to 30 l/s, respectively. Only for the daily flow data (Figure 2-8d), the threshold estimated by the SE method was smaller than those estimated from the Normality of Differences tests and relatively close to the threshold estimated by ATS (Figure 2-7d). For hourly flow data (Figure 2-7b and Figure 2-8b), ATS and Pearson's chi square test (for Normality of Differences) provided almost identical estimates, while all other methods suggested much larger thresholds. Noticeably, the hourly thresholds estimated by the SE method and the Shapiro-Francia test are very close at 44.68 l/s and 45.33 l/s, respectively (Figure 2-8b), but result in considerably different shape parameters (Table 2-1). Figure 2-7b reveals hourly thresholds to be in the region where the shape characteristics show large fluctuations due to the small sample size that results in an inefficient fit of the GPD and likely spurious estimates of the shape parameter.

The performance of the Normality of Differences method depended greatly on both the given normality test and on data resolution. For the 15-minute flow data, all normality tests provided relatively similar threshold selections (Figure 2-8a), which was not the case for the hourly and 6-hourly flow data (Figure 2-8b and Figure 2-8c). For the daily flow data (Figure 2-8d), thresholds were estimated too large and consequently result in too few values for efficient statistical inference. In general, the smaller the selected threshold, given that the excesses are satisfactorily modelled by the GPD, the lower the uncertainty and consequently the lower the variance in the parameter estimates due to larger sample sizes.

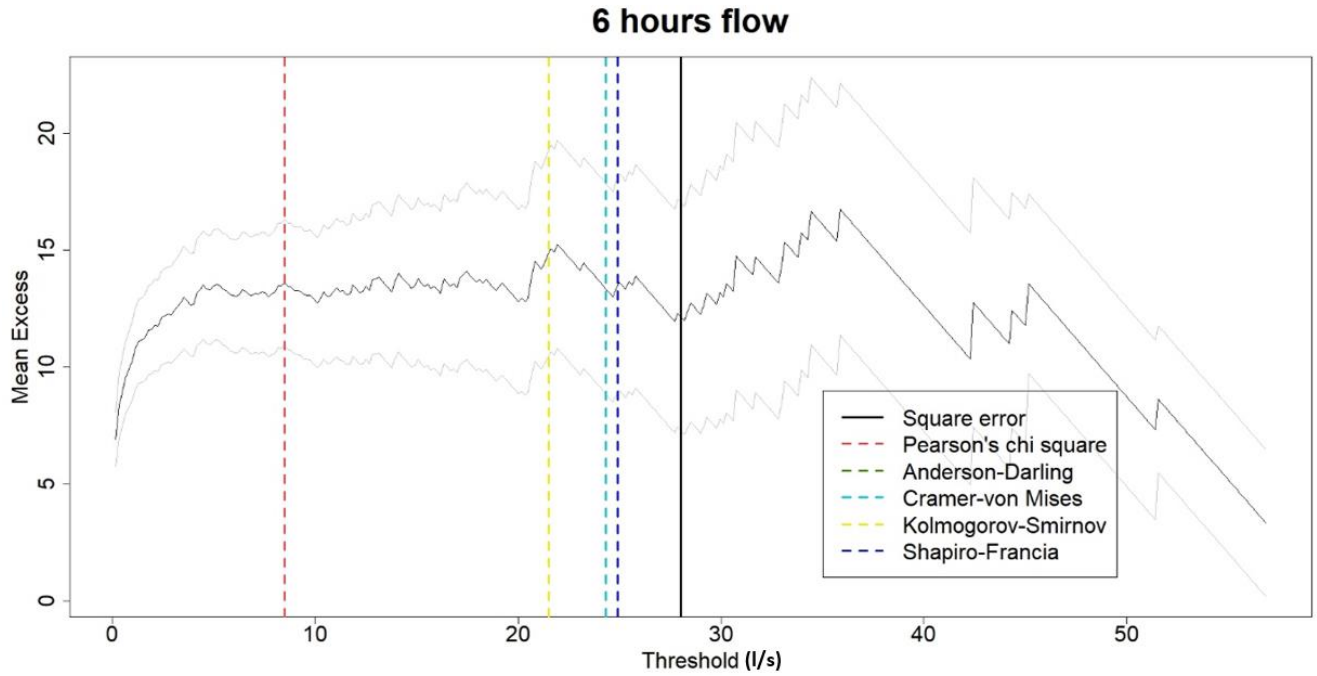
a)


15 minutes flow
Mean Excess
Threshold (l/s)
Square error
Pearson's chi square
Anderson-Darling
Cramer-von Mises
Kolmogorov-Smirnov
Shapiro-Francia

b)


Hourly flow
Mean Excess
Threshold (l/s)
Square error
Pearson's chi square
Anderson-Darling
Cramer-von Mises
Kolmogorov-Smirnov
Shapiro-Francia

c)



d)

Figure 2-8: MRL plots: Mean excesses and their 95% confidence intervals plotted against threshold for the a) 15 minutes, b) hourly, c) 6 hourly and d) daily flow data. The threshold selected using the SE method is shown by the vertical solid line and the thresholds selected by the Normality of Differences tests are shown by the dashed vertical lines.

2.5.2.4 Parameter and fit comparisons

In summary, the estimated shape parameters showed little consistency across the four data resolutions and across the threshold selection techniques investigated (Table 2-1). The 15-minute extreme flows are characterized by: (i) an exponential tail (Pearson's chi square, Anderson Darling and Kolmogorov-Smirnov tests) as the shape parameter takes values close to zero, (ii) heavy tails (SE method, Shapiro-Francia and Cramer-von Mises tests) and (iii) short tails ($\xi < 0$) (ATS method). ATS and Normality of Differences methods resulted in short tail distributions for both the hourly and 6-hourly flow data, whereas the SE method resulted in a heavier tail, similar to that found across all flow data scales. The ATS and the SE methods provided heavy tails for the daily flow, and the Normality of Differences tests tended to short tails.

Table 2-1: Estimated thresholds and shape parameters for four flow resolutions and three core threshold selection methods.

		ATS	SE	Normality of Differences tests				
				Pearson's chi square	Anderson-Darling	Cramer-von Mises	Kolmogorov-Smirnov	Shapiro-Francia
15 mins	Threshold	22.2	46.8	39.7	51.8	45.5	42.6	53.5
	Shape Parameter	-0.14	0.33	0.01	0.07	0.26	0.06	0.10
Hourly	Threshold	9.7	44.7	9.6	66.9	70.7	80.1	45.3
	Shape Parameter	-0.09	0.17	-0.09	-0.58	-0.44	-0.48	-0.35
6 hours	Threshold	6.6	28.1	8.5	24.3	24.3	21.5	24.9
	Shape Parameter	-0.01	0.20	-0.05	-0.23	-0.23	-0.34	-0.23
Daily	Threshold	3.1	5.6	17.3	17.8	18.4	19.3	17.1
	Shape Parameter	0.17	0.22	-0.17	-0.10	-0.08	0.10	-0.20

Table 2-2: MSE between the empirical and theoretical quantiles for different threshold selection methods at four flow resolutions.

MSE	ATS	SE	Normality of Differences tests				
			Pearson's chi square	Anderson-Darling	Cramer-von Mises	Kolmogorov-Smirnov	Shapiro-Francia
15 mins	252.4	8248.8	123.7	2157.8	6034.9	1242.3	2828.2
Hourly	130.9	2654.1	24.1	14.5	13.6	10.5	28.0
6 hourly	72.1	150.8	61.0	34.0	34.0	12.7	34.8
Daily	38.2	81.9	8.3	10.7	12.6	32.4	7.6

The MSE (Table 2-2) seems to be an inappropriate diagnostic for deviations between very large theoretical and empirical quantiles as it depends greatly on the shape parameter. Peak flows with very short finite tails will show minimum MSEs, which increase by orders of magnitude as the shape parameter increases. Conversely, the NRMSE does provide a comparative diagnostic since it is normalized by accounting for very large values that are associated with heavy tails. Thus, NRMSE values are reported in Table 2-3 where compared to the SE and Normality of Differences methods, this study's ATS method gives the smallest NRMSE for flow data of any resolution, except for the Normality of Differences tests for the hourly flow.

Table 2-3: NRMSE between the empirical and theoretical quantiles for different threshold selection methods at four flow resolutions.

NRMSE	ATS	SE	Normality of Differences tests				
			Pearson's chi square	Anderson-Darling	Cramer-von Mises	Kolmogorov-Smirnov	Shapiro-Francia
15 mins	102.6	1017.9	308.0	571.6	866.6	391.4	697.5
Hourly	38.8	244.4	37.7	30.9	29.9	38.2	27.0
6 hourly	51.8	184.2	67.6	87.4	87.4	53.4	88.5
Daily	44.5	69.3	52.6	59.5	72.0	115.3	50.2

The relative index of agreement (Figure 2-9) is also an efficient measure of proximity between observed and simulated peak flows (Krause et al., 2005). For this diagnostic, the GPD was

consistently best fitted to empirical peak flows at all scales when their thresholds were chosen using this study's ATS method. Here, the SE method was the poorest method, especially at the 15-minute data scale. Interestingly, results at the hourly scale behaved very differently to those found at the three other scales. We speculate that this was likely due to the hourly data being at, or close to, the natural water run-off integration rate to the sub-catchment's water flume following a rainfall event (see Discussion).

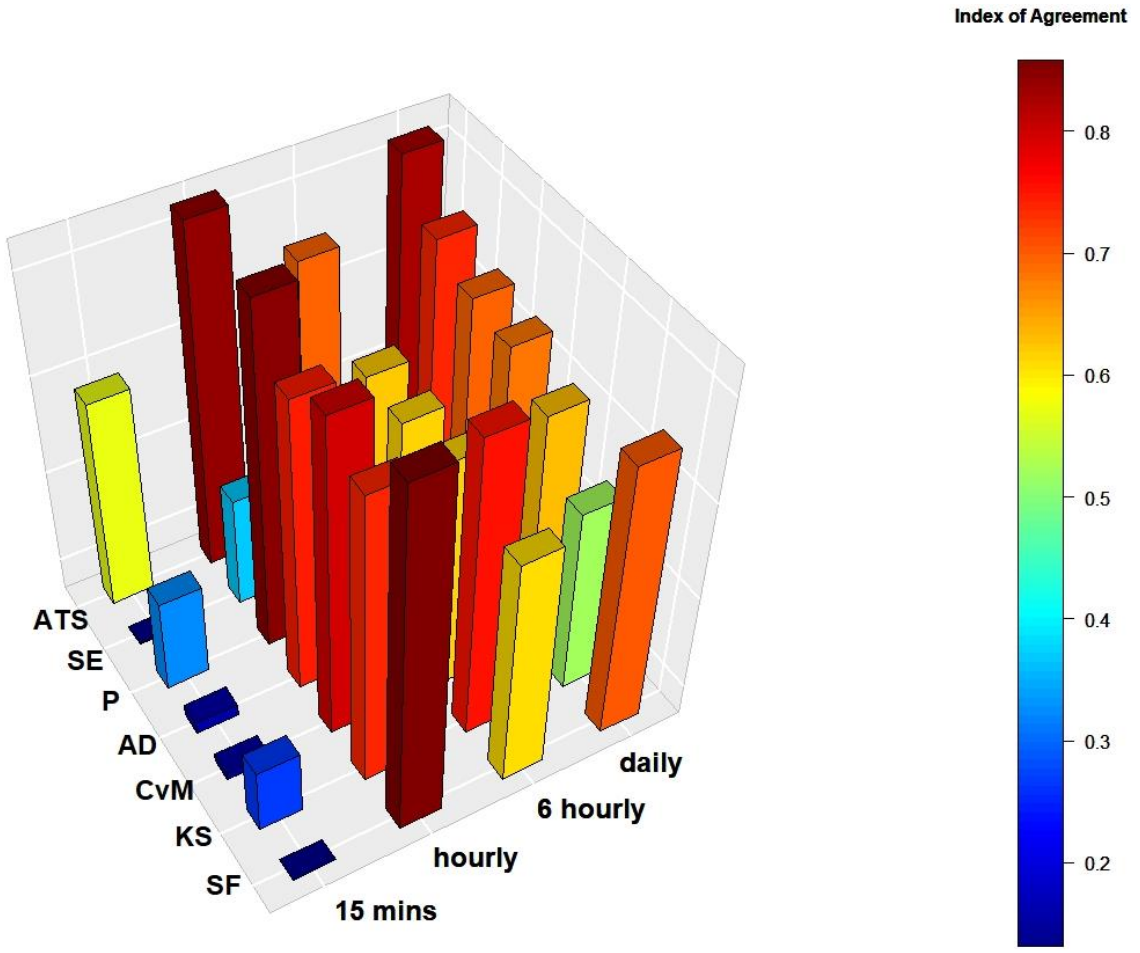


Figure 2-9: Index of agreement between theoretical and empirical peak flow of different resolutions. The value of one corresponds to a perfect match and the value of zero represents no agreement at all. The threshold selection methods are Automated Threshold Stability (ATS), Square Error (SE) and the various tests of the Normality of Differences method, the Pearson's chi-square (P), Anderson-Darling (AD), Cramer-von Mises (CvM), Kolmogorov-Smirnov (KS) and Shapiro-Francia (SF).

Figure 2-10 presents the Q-Q plots of the 15-minute extreme flows for the threshold selection methods that gave the smallest (ATS) and the largest (SE) NRMSE values (Table 2-3). The Q-Q plots show that an over-estimated threshold results in a sample size that can be too small for efficient statistical inference and results in increased uncertainty. The Q-Q plots also emphasis the superiority of this study's ATS method given its Q-Q plot falls relatively close to the 45° line.

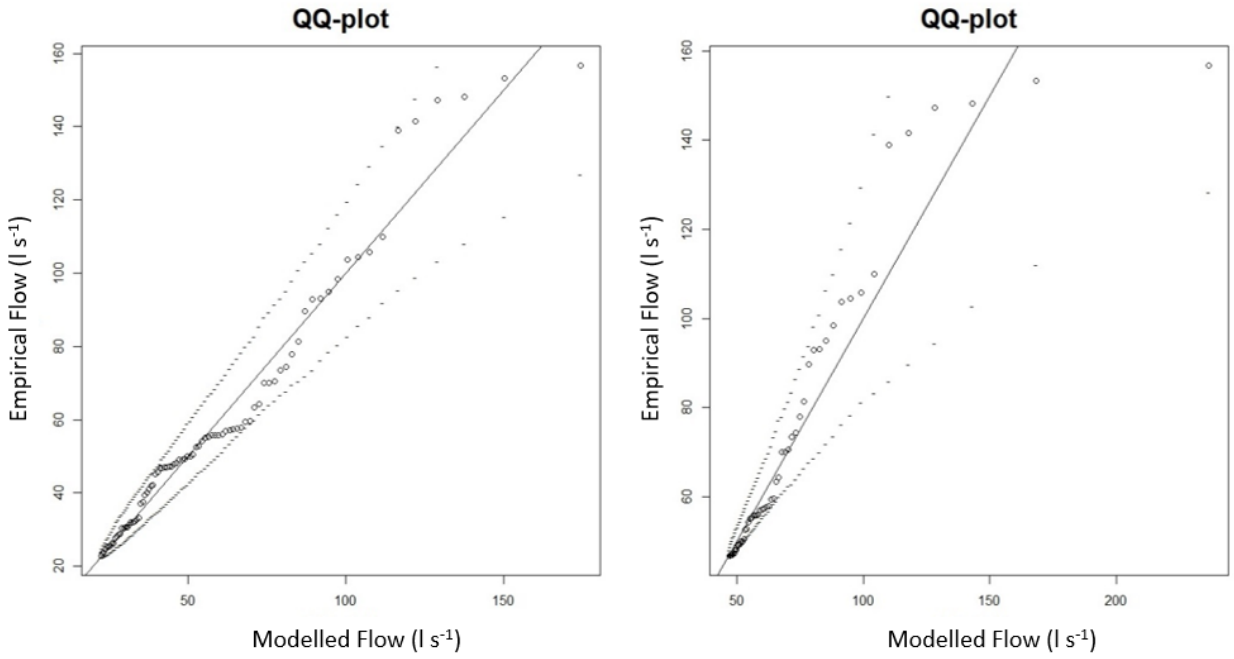


Figure 2-10: Q-Q plots of the 15-minute peak flows estimated by the ATS (left) and SE (right) methods. The solid line is the model, the dashed lines correspond to 95% confidence intervals and the points depict the measured peak flows. The number of points is a function of the selected threshold.

Clear differences are observed in the estimated Return Level / Return Period plots for the ATS and Normality of Difference (Kolmogorov-Smirnov test only) methods (Figure 2-11). For example, ATS suggests that a daily peak of 25 l s^{-1} are observed once in 50 years on average, whereas according to Normality of Difference using the Kolmogorov-Smirnov test, peaks of such magnitude should be expected every 5 years. This indicates that the combined effects

of data scale, the GPD estimator and the threshold selection method - each have a significant impact on the characteristics of the final model that attempts to explain the flow process with the consideration of extremes. This is critically important in cases where reliably informed actions need to be taken or infrastructure needs to be built to mitigate the impacts of future peak flows and likely flood events.

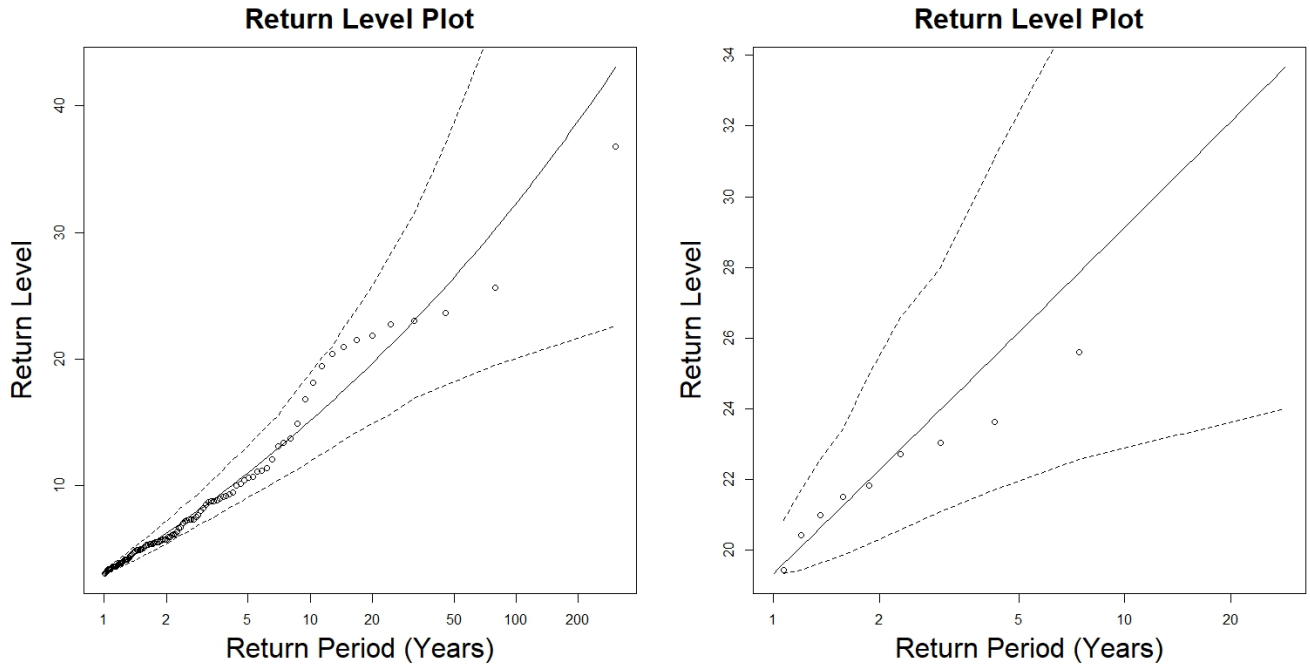


Figure 2-11: Return level plots of the daily peak flows estimated by the ATS (left) and Normality of Difference Kolmogorov-Smirnov (right) methods. The solid line is the model, the dashed lines correspond to 95% confidence intervals and the points depict the measured peak flows. The number of points is a function of the selected threshold.

2.6 Discussion

In agreement with previous studies (e.g. Bermudez & Kotz, 2010; Engeland et al., 2004), we found that the performance of the GPD parameter estimators examined through a Monte Carlo experiment, depended significantly on the sample size and the value of the shape parameter. The MLE/MPLE, PWMU/PWMB and the LME were consistently the most unbiased

and precise estimators and so we chose only from this group in our subsequent analyses. More specifically, for the application of the SE and AST threshold selection methods, a different GPD estimator was used each time according to its strengths. For example, the LME was preferred for positive shape parameters and large sample size.

This study's Automated Threshold Stability (ATS) method was tested against existing SE and Normality of Differences methods. Methods were applied to flow discharge measurements of 15-minute resolution, as well as to the same data aggregated to coarser resolutions of hourly, 6-hourly and daily, to examine scale effects. The Normality of Differences method depended on the normality test applied and resulted in short, exponential and heavy tailed distributions even at the same scale (e.g. shape parameters of $\xi = -0.2$ for the daily flow according to Shapiro-Francia and $\xi = 0.1$ according to the Kolmogorov-Smirnov test). Similar results for the value of the shape parameter were obtained from the ATS method, unlike the SE method which always resulted in positive ξ.

Threshold stability plots were discussed in Scarrott and MacDonald (2012) and Solari and Losada (2012), but these studies did not perform an analytical approximation, as done here with ATS, although Langousis et al. (2016) suggested an automated technique based on the assumption of linearity of the MRL plot and applied it to rainfall data. Our proposed ATS method provided more robust estimates of the threshold compared to: (a) the SE method as it was less sensitive to the resolution of the data and (b) the Normality of Differences method as it was less sensitive to the sample size of the threshold candidates. It also resulted in the smallest errors and the largest agreement indices between the simulated and the empirical quantiles.

Specific to the case study, error and agreement indices indicated that the GPD provided the best fit to the hourly peak flow data relative to 15-minute, 6-hourly and daily peak flow data. For all the applied threshold selection methods, the modelled peak flow at the hourly resolution was consistently the closest to the empirical one, compared to three other scales. These results cannot be attributed to the value of the shape parameter (e.g. short finite tails result in greater agreement between theoretical and empirical quantiles) since the SE method gives a positive ξ. An inspection of the plots and a comparison across various scales does not reveal any pattern that would justify this behavior. A possible explanation could be that the hourly peak flow best captures the signal of the process and integrates more efficiently the way the 6.84 ha sub-catchment (of two pasture fields) transforms intense rainfall into high discharge. It should be noted that the data aggregation was not done at equal intervals. For example, the hourly flow resulted from averaging four 15-minute measurements, whereas the 6-hourly and the daily flow are the averages of 24 and 96 observations, respectively. This does not affect the results but should be borne in mind when interpreting the plots.

An advantage of using fine resolution flow data is that they result in larger sample sizes that can make the statistical inference more efficient even for records of short periods for which a GEV/AM extreme value methodology is not applicable. However, this study showed that for data of the same resolution, the value of the GPD shape parameter varies according to the selected thresholds. This has serious practical implications since the models are commonly extrapolated beyond observed values for forecasting and engineering design purposes to mitigate against future flooding. On one hand, an under-estimated threshold and shape parameter of the extreme flow can result in failure of hydrological infrastructure (e.g. dams,

flood protection works) due to higher peak flows than expected. On the other hand, over-estimation of the high flows can lead to over-pricing and mis-use of resources.

2.7 Conclusions

In this study, we examined the effect of statistical estimators, data resolution, and threshold selection on fitting the Generalized Pareto distribution to peak hydrological flows that resulted from the 'Peaks Over Threshold' method. Through a simulation study, the performance of the estimators depended greatly on the sample size and the shape parameter where the only most accurate and unbiased estimators were used for the selection of thresholds in subsequent empirical evaluations. Here an automated threshold selection method based on the stability of the shape and modified scale parameters was empirically demonstrated to provide more robust estimates compared to two commonly applied alternatives. The proposed method provided the smallest error and the greatest agreement indices between the empirical and theoretical quantiles across all the scales of the case study flow data. showed

The study results can be generalized to similar water monitoring schemes for improved characterization of likely flood events. However, the study highlights that the combined effect of data scale, threshold selection method and statistical estimator, significantly affects the shape parameter and, as a consequence, the nature of the Generalized Pareto distribution. Such linked effects need to be acknowledged and assessed as they have clear implications for the reliable forecasting of extreme flow events, and the consequences thereof.

Authors contribution statement

Stelian Curceac 80%: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing.

Peter M. Atkinson 5%: Conceptualization, Writing - review & editing, Supervision, Funding acquisition.

Alice Milne 5%: Conceptualization, Writing - review & editing, Supervision, Funding acquisition.

Lianhai Wu 5%: Writing - review & editing, Supervision, Funding acquisition.

Paul Harris 5%: Conceptualization, Data curation, Writing - review & editing, Supervision, Funding acquisition.

Acknowledgements

Rothamsted Research receives grant aided support from the Biotechnology and Biological Sciences Research Council (BBSRC) of the United Kingdom. This research was funded by Rothamsted Research and Lancaster Environment Centre, the BBSRC Institute Strategic Programme (ISP) grant, "Soils to Nutrition" (S2N) grant numbers BBS/E/C/000I0320, BBS/E/C/000I0330 and the BBSRC National Capability grant for the North Wyke Farm Platform grant number BBS/E/C/000J0100.

Declaration of interest

The authors declare no potential conflict of interest associated with this research.

Software and data availability

The statistical software (R Core Team, 2017) and all North Wyke Farm Platform data sets (https://www.rothamsted.ac.uk/north-wyke-farm-platform) are freely available.

References

Ashkar, F. and Tatsambon, C. N. (2007). Revisiting some estimation methods for the generalized Pareto distribution. Journal of Hydrology, 346, 136-143.

Bates, B. C., Kundzewicz, Z. W., Wu, S. and Palutikof, J. P. (2008). Climate Change and Water. Technical Paper of the Intergovernmental Panel on Climate Change, IPCC Secretariat, Geneva, 210 pp.

Beguería, S. (2005). Uncertainties in Partial Duration Series Modelling of Extremes Related to the Choice of the Threshold Value. Journal of Hydrology, 303(1), 215-230.

Behrens, C. N., Lopes, H. F. and Gamerman, D. (2004). Bayesian Analysis of Extreme Events with Threshold Estimation. Statistical Modelling, 4(3), 227-244.

Beirlant, J., Dierckx, G. and Guillou, A. (2005). Estimation of the Extreme-Value Index and Generalized Quantile Plots. Bernoulli, 11(6), 949-970.

Beirlant, J., Joossens, E. and Segers, J. (2005). Unbiased tail estimation by an extension of the generalized Pareto distribution. CentER Discussion Paper, Vol. 2005–112, Tilburg: Econometrics.

Beirlant, J., Vynckier, P. and Teugels, J. L. (1996). Tail Index Estimation, Pareto Quantile Plots, and Regression Diagnostics. Journal of the American Statistical Association, 91(436), 1659-1667.

Beirlant, J., de Wet, T. and Goegebeur, Y. (2006). A Goodness-of-Fit Statistic for Pareto-Type Behaviour. Journal of Computational and Applied Mathematics, Special Issue: Jef Teugels, 186(1), 99-116.

Bommier, E. (2014). Peaks-over-threshold modelling of environmental data (Technical report). U.U.D.M. Project Report, 2014:33.

Bouraoui, F., Grizzetti, B., Granlund, K., Rekolainen, S. and Bidoglio, G. (2004). Impact of Climate Change on the Water Cycle and Nutrient Losses in a Finnish Catchment, Climatic Change, 66(1–2), 109-126.

Brodin, E. and Rootzén, H. (2009). Univariate and Bivariate GPD Methods for Predicting Extreme Wind Storm Losses. Insurance: Mathematics and Economics, 44(3), 345-356.

Choulakian, V. and Stephens, M. A. (2001). Goodness-of-Fit Tests for the Generalized Pareto Distribution. Technometrics, 43(4), 478-484.

Claps, P. and F. Laio. (2003). Can Continuous Streamflow Data Support Flood Frequency Analysis? An Alternative to the Partial Duration Series Approach. Water Resources Research, 39(8).

Clarke, M. L. & Rendell, H. M. (2006). Hindcasting Extreme Events: The Occurrence and Expression of Damaging Floods and Landslides in Southern Italy. Land Degradation & Development, 17(4), 365-380.

Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer, London, UK.

Coles, S. and Dixon, M. J. (1999). Likelihood-Based Inference for Extreme Value Models, Extremes, 2(1), 5-23.

Cunnane, C. (1979). A Note on the Poisson Assumption in Partial Duration Series Models. Water Resources Research, 15(2), 489-494.

Danielsson, J., de Haan, L., Peng, L. and de Vries, C. G. (2001). Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation. Journal of Multivariate Analysis, 76(2), 226-248.

Das, B. and Ghosh, S. (2013). Weak limits for exploratory plots in the analysis of extremes. Bernoulli, 19(1), 308-343

Davison, A. C. and Smith, R. L. (1990). Models for Exceedances over High Thresholds, Journal of the Royal Statistical Society, Series B (Methodological), 52(3), 393-442.

Deidda, R. (2010). A Multiple Threshold Method for Fitting the Generalized Pareto Distribution to Rainfall Time Series. Hydrology and Earth System Sciences, 14(12), 2559-2575.

Deidda, R. and Puliga, M. (2006). Sensitivity of Goodness-of-Fit Statistics to Rainfall Data Rounding Off. Physics and Chemistry of the Earth, 31(18). 1240-1251.

Dekkers, A. L. M. and De Haan, L. (1989). On the Estimation of the Extreme-Value Index and Large Quantile Estimation, The Annals of Statistics, 17(4), 1795-1832.

Durocher, M., Zadeh, S. M., Burn, D. H. and Ashkar, F. (2018). Comparison of Automatic Procedures for Selecting Flood Peaks over Threshold Based on Goodness-of-Fit Tests. Hydrological Processes, 32(18), 2874–2887.

Eastoe, E. F. and Tawn, J. A. (2010). Statistical Models for Overdispersion in the Frequency of Peaks over Threshold Data for a Flow Series. Water Resources Research, 46(2).

Engeland, K., Hisdal, H. and Frigessi, A. (2004). Practical Extreme Value Modelling of Hydrological Floods and Droughts: A Case Study. Extremes, 7(1), 5–30.

Ferguson, T. S., Genest, C. and Hallin, M. (2000). Kendall's Tau for Serial Dependence. Canadian Journal of Statistics, 28(3), 587–604.

Field, C. B., Barros, V., Stocker, T. F. and Dahe, Q. (2012). Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change, Cambridge, Cambridge University Press.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample, Proc. Cambridge Philos. Soc., 24(2), 180-190.

Gharib, A., Davies, E. G. R., Goss, G. G. and Faramarzi, M. (2017). Assessment of the Combined Effects of Threshold Selection and Parameter Estimation of Generalized Pareto Distribution with Applications to Flood Frequency Analysis, Water, 9(9), 692.

Goegebeur, Y., Beirlant, J. and de Wet, T. (2008). Linking Pareto-Tail Kernel Goodness-of-Fit Statistics with Tail Index at Optimal Threshold and Second Order Estimation., REVSTAT-Statistical Journal, 6(1), 51-69.

Greenwood, J. A., Landwehr, J. M., Matalas N. C. and Wallis, J. R. (1979). Probability Weighted Moments: Definition and Relation to Parameters of Several Distributions Expressable in Inverse Form. Water Resources Research, 15(5), 1049-1054.

Hall, P. (1990). Using the Bootstrap to Estimate Mean Squared Error and Select Smoothing Parameter in Nonparametric Problems. Journal of Multivariate Analysis, 32(2), 177-203.

Hill, B. M. (1975). A Simple General Approach to Inference About the Tail of a Distribution. The Annals of Statistics, 3(5), 1163-1174.

Hosking, J. R. M. and Wallis, J. R. (1987). Parameter and Quantile Estimation for the Generalized Pareto Distribution. Technometrics, 29(3), 339-349.

Jenkinson, A. F. (1955). The Frequency Distribution of the Annual Maximum (or Minimum) Values of Meteorological Elements. Quarterly Journal of the Royal Meteorological Society, 81(348), 158-171.

Josse, J. and Husson, F. (2013). Handling Missing Values in Exploratory Multivariate Data Analysis Methods. Journal de La Société Française de Statistique, 153(2), 79–99.

Krause, P., Boyle, D. P. and Bäse, F. (2005). Comparison of Different Efficiency Criteria for Hydrological Model Assessment. Advances in Geosciences, 5, 89–97.

Kundzewicz, Z. W., Mata, L. J., Arnell, N. W., Doll, P., Kabat, P., Jimenez, B. et al. (2007). Freshwater Resources and Their Management. In Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by M. L. Parry, O. F. Canziani, J. P. Palutikof, P. J. van der Linden, and C. E. Hanson, 173–210. Cambridge University Press.

Landwehr, J. M., Matalas, N. C. and Wallis, J. R. (1979). Probability Weighted Moments Compared with Some Traditional Techniques in Estimating Gumbel Parameters and Quantiles. Water Resources Research, 15(5), 1055-1064.

Lang, M., Ouarda, T. B. M. J. and Bobée, B. (1999). Towards Operational Guidelines for Over-Threshold Modeling. Journal of Hydrology, 225(3), 103-117.

Langousis, A., Mamalakis, A., Puliga, M. & Deidda, R. (2016). Threshold Detection for the Generalized Pareto Distribution: Review of Representative Methods and Application to the NOAA NCDC Daily Rainfall Database. Water Resources Research, 52(4), 2659–2681.

Ledford, A. W. and Tawn, J. A. (2003). Diagnostics for Dependence within Time Series Extremes. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 65(2), 521–543.

Liang, B, Shao, Z., Li, H., Shao, M. and Lee, D. (2019). An Automated Threshold Selection Method Based on the Characteristic of Extrapolated Significant Wave Heights. Coastal Engineering, 144, 22-32.

Liu, Y., Li, Y., Harris, P., Cardenas, L. M., Dunn, R. M., Sint, H., Murray, P. J., Lee, M. R. F. and Wu, L. (2018). Modelling Field Scale Spatial Variation in Water Run-off, Soil Moisture, N2O Emissions and Herbage Biomass of a Grazed Pasture Using the SPACSYS Model. Geoderma, 315, 49-58.

Luceño, A. (2006). Fitting the Generalized Pareto Distribution to Data Using Maximum Goodness-of-Fit Estimators. Computational Statistics & Data Analysis, 51(2), 904-917.

Mackay, E. B. L., Challenor, P. G. and Bahaj, A. S. (2011). A Comparison of Estimators for the Generalised Pareto Distribution. Ocean Engineering, 38(11), 1338-1346.

Madsen, H., Rasmussen, P. F. and Rosbjerg, D. (1997). Comparison of Annual Maximum Series and Partial Duration Series Methods for Modeling Extreme Hydrologic Events: 1. At-Site Modeling. Water Resources Research, 33(4), 747–757.

Millenniu Ecosystem Assessment, 2005. Millenniu Ecosystem Assessment (MA), Ecosystems and Human Well-being: Synthesis. Island Press, Washington, DC.

Moore, D. S. (1986). Tests of Chi-Squared Type Goodness of Fit Techniques. Marcel Dekker, New York.

Orr, R. J., Murray, P. J., Eyles, C. J., Blackwell, M. S. A., Cardenas, L. M., Collins, A. L. et al. (2016). The North Wyke Farm Platform: effect of temperate grassland farming systems on soil moisture contents, runoff and associated water quality dynamics, European Journal of Soil Science, 67, 374–385.

de Zea Bermudez, P. and Kotz, S. (2010). Parameter Estimation of the Generalized Pareto Distribution, Part I. Journal of Statistical Planning and Inference, 140(6), 1353-1373.

Pickands, J. (1975). Statistical Inference Using Extreme Order Statistics. The Annals of Statistics, 3(1), 119-131.

Prescott, P. and Walden, A. T. (1980). Maximum Likelihood Estimation of the Parameters of the Generalized Extreme-Value Distribution. Biometrika, 67(3), 723-724.

Prescott, P. and Walden, A. T. (1983). Maximum Likelihood Estimation of the Parameters of the Three-Parameter Generalized Extreme-Value Distribution from Censored Samples. Journal of Statistical Computation and Simulation, 16(3–4), 241-250.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reiss, R. D. and Thomas, M. (2007). Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields. 3rd edition. Birkhäuser Basel.

Scarrott, C. and MacDonald, A. (2012). A Review of Extreme Value Threshold Es-Timation and Uncertainty Quantification. REVSTAT–Statistical Journal, 10(1), 33–60.

Segers, J. (2005). Generalized Pickands Estimators for the Extreme Value Index. Journal of Statistical Planning and Inference, 128(2), 381-396.

Sheta, A. F. and El-Sherif, M. S. (1999). Optimal Prediction of the Nile River Flow Using Neural Networks. International Joint Conference on Neural Networks. Proceedings, 5, 3438-3441.

Sigauke, C. and Bere, A. (2017). Modelling Non-Stationary Time Series Using a Peaks over Threshold Distribution with Time Varying Covariates and Threshold: An Application to Peak Electricity Demand. Energy, 119, 152-166.

Smith, R. L. (1985). Maximum Likelihood Estimation in a Class of Nonregular Cases. Biometrika, 72(1), 67-90.

Solari, S. and Losada, M. A. (2012). A Unified Statistical Model for Hydrological Variables Including the Selection of Threshold for the Peak over Threshold Method. Water Resources Research, 48(10).

Solari, S., Egüen, M., Polo, M. J. and Losada, M. A. (2017). Peaks Over Threshold (POT): A Methodology for Automatic Threshold Estimation Using Goodness of Fit p-Value. Water Resources Research, 53(4), 2833–2849.

Takahashi, T., Harris, P., Blackwell, M. S. A., Cardenas, L. M., Collins, A. L., Dungait, J. A. J. et al. (2018). Roles of Instrumented Farm-Scale Trials in Trade-off Assessments of Pasture-Based Ruminant Production Systems. Animal, 12 (8), 1766-1776.

Tanaka, S. and Takara, K. (2002). A Study on Threshold Selection in POT Analysis of Extreme Floods, Extremes of the Extremes, Extraordinary Floods, IAHS Publ, 271, 299-304.

Thibault, K. M. & Brown, J. H. (2008). Impact of an Extreme Climatic Event on Community Assembly. Proceedings of the National Academy of Sciences, 105(9), 3410-3415.

Thode, H. C. (2002). Testing For Normality. CRC Press.

Thompson, P., Cai, Y., Reeve, D. and Stander, J. (2009). Automated Threshold Selection Methods for Extreme Wave Analysis. Coastal Engineering, 56(10), 1013-1021.

Todorovic, P. (1978). Stochastic Models of Floods. Water Resources Research, 14(2), 345–356.

Turan, M. E., and Yurdusev, M. A. (2009). River Flow Estimation from Upstream Flow Records by Artificial Intelligence Methods. Journal of Hydrology, 369(1), 71-77.

Willmott, C. J. (1981). On the Validation of Models. Physical Geography, 2(2), 184-194.

Yang, X., Zhang, J. and Ren, W. X. (2018). Threshold Selection for Extreme Value Estimation of Vehicle Load Effect on Bridges. International Journal of Distributed Sensor Networks, 14(2).

Yun, S. (2002). On a Generalized Pickands Estimator of the Extreme Value Index. Journal of Statistical Planning and Inference, 102(2), 389-409.

de Zea Bermudez, P. and Kotz, S. (2010). Parameter Estimation of the Generalized Pareto Distribution, Part I. Journal of Statistical Planning and Inference, 140(6), 1353–1373.

Zhang, J. (2007). Likelihood Moment Estimation for the Generalized Pareto Distribution. Australian & New Zealand Journal of Statistics, 49(1), 69-77.

Zoglat, A., EL Adlouni, S., Badaoui, F., Amar A. & Okou, C. G. (2014). Managing Hydrological Risks with Extreme Modeling: Application of Peaks over Threshold Model to the Loukkos Watershed, Morocco, Journal of Hydrologic Engineering, 19(9), 05014010.

3. Effects of data temporal resolution on the simulation of water flux extremes using a process-based model at the grassland field scale

Lianhai Wu [a,*], Stelian Curceac [a], Peter M. Atkinson [b,c,d], Alice Milne [e], Paul Harris [a]

[a] Rothamsted Research, Department of Sustainable Agriculture Sciences, North Wyke EX20 2SB, Devon, UK.

[b] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK.

[c] Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

[d] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China.

[e] Rothamsted Research, Department of Sustainable Agriculture Sciences, Harpenden AL5 2JQ, UK

* Corresponding author: Lianhai Wu

Email: lianhai.wu@rothamsted.ac.uk

Published in Agricultural Water Management

3.1 Abstract

Projected changes to rainfall patterns may exacerbate existing risks posed by flooding. Furthermore, increased surface and sub-surface runoff from agricultural land increases pollution through nutrient losses. Agricultural systems are complex because they are managed in individual fields, and it is impractical to provide resources to monitor their water fluxes. In this respect, modelling provides an inexpensive tool for simulating fluxes. Due to data availability and the dynamics of the processes at the field-scale, a daily time-step is used routinely However, it was hypothesized that a finer time-step will provide more accurate identification of peak fluxes. To investigate this, a 15-minute water flux dataset from April 2013 to February 2016 from a pasture within a monitored grassland research farm was up-scaled to hourly, 6-hourly and daily data; and a daily time-step process-based model was adapted to provide corresponding down-scaled simulations at 15-minute, hourly and 6-hourly resolution (in addition to its usual daily output). Analyses were conducted with respect to model performance for: (a) each of the four data resolutions, separately (15-minute measured versus 15-minute simulated; hourly measured versus hourly simulated; etc.); and (b) at the daily resolution only, where 15-minute, hourly and 6-hourly simulations were each aggregated to the daily scale. Comparison between measured and simulated fluxes at the four resolutions (unaggregated approach) revealed that hourly simulations provided the highest percentage of correctly identified water flux peaks. Conversely, aggregating to the daily scale using either 15-minute or hourly simulations increased accuracy, both in prediction of general trends and identification of peak fluxes. The improved identification of extremes resulted in 9 out of 11 peak flow events being correctly identified with only 2 false positives, compared with 5 peaks being identified with 4 false positives of the usual daily simulations. Increased

peak flow detection accuracy has the potential to provide clear field management benefits in reducing nutrient losses to water.

Key words: SPACSYS; extreme flows; North Wyke Farm Platform; scale effects; grassland;

3.2 Introduction

Flooding in the UK puts more than 5 million people in 2.4 million properties at risk each year (Environment Agency, 2009). Projected changes to rainfall patterns (Watts and Anderson, 2016) may exacerbate the existing risks posed by flooding. Flash flooding or surface water flooding is one of the most common types of flooding in the UK. It is defined as those flood events where the rise in water is either during or within a few hours of the rainfall that produces the rise exceeding the capacity of a river or creating floods even in locations far from water bodies. The utilised agricultural area, of which almost 60% is permanent grassland, covers 71% of the total land of the UK (Department for Environment, Food and Rural Affairs, 2019). Water fluxes or surface runoff generated from agricultural land can contribute significantly to local floods and nutrient losses that cause water pollution. Flooding of farmland is likely to become more frequent in some areas under projected climate change (Brown et al., 2016), although intriguingly, studies have found increases in precipitation extremes do not necessarily mean increases in flood magnitude, due to decreased soil moisture at storm onset during the dry months and reduced storm durations (Sharma et al., 2018; Wasko et al., 2019). Further, the combined effect of increased rainfall intensity, slope, land use and antecedent soil moisture can accelerate soil erosion (Ziadat and Taimeh, 2013), leading to the loss of valuable topsoil and the pollution of watercourses (Morison and Matthews, 2016).

Accurate forecasting of water runoff (or water fluxes) from agricultural land is, therefore, not only a vital component of flood early-warning systems, but also for associated management strategies for nutrient loss and water pollution. Water fluxes from the soil surface are controlled by soil properties. Long-term hydrological studies have shown that sandy Alfisols can generate higher runoff compared to clayey Vertisols (Pathak et al., 2013), and a greater risk of flooding on clay soils has been reported (Charlton et al., 2010). The wetness of the soil before a precipitation event (Merz and Plate, 1997) and soil compaction also affect water fluxes. Farm machinery and livestock (Adimassu et al., 2019; Alaoui et al., 2018; Newell Price et al., 2012) can cause serious compaction and so exacerbate runoff risk. Natural events, particularly long and intense precipitation events (Archer and Fowler, 2018), and land cover variation (Dadson et al., 2017; Keesstra et al., 2018) also affect flux.

Agricultural systems are complex because they are generally managed at the field scale and each field has its own unique set of soil conditions and topography. Monitoring water surface fluxes in fields is costly both in time and financially. In this respect, modelling provides an effective tool for simulating or forecasting water fluxes. The SPACSYS model (Wu et al., 2007) is one such process-based model. It is a field scale and weather-driven dynamic simulation model. Since it was first published in 2007, it has been developed to provide added functionality (Bingham and Wu, 2011; Liu et al., 2013; Wu et al., 2019; Wu et al., 2015). The model can simulate the interactions of soil carbon (C), nitrogen (N) and phosphorus (P), plant growth and development, water re-distribution and heat transformation in agricultural fields. The model has been used to investigate several issues including resource use efficiency by crops (Wu et al., 2009), greenhouse gas (GHG) emissions (Abalos et al., 2016; Perego et al.,

2016), the responses of cropping/grassland systems to environmental change (Wu et al., 2016) and the forecasting of crop yield and stocks of C and nutrients (Zhang et al., 2016) under various climatic and soil conditions.

The SPACSYS model has been developed to investigate not only temporal dynamics, but also within-field spatial variation in processes such as water runoff, using a linked, grid-based approach (grid-to-grid) (Liu et al., 2018). As in all previous implementations of SPACSYS, and common to many agriculture-focused models (Ahuja et al., 2002), a daily time-step has been used. However, model predictions of water flux did not increase in accuracy when considering a within-field grid connectivity approach (Liu et al., 2018). We hypothesise, that a finer time-step might provide this improvement instead; not only in the grid-to-grid model, but also in the (non-grid-to-grid) standard model, as investigated here. Although not demonstrated within this study, increasing the accuracy of water flux simulations should implicitly increase the accuracy of associated SPACSYS simulations, such as those for nutrient loss that use predicted water flux in their calculation.

For our case study, we used measured 15-minute water flux data from one field (or sub-catchment) of the North Wyke Farm Platform (NWFP). The NWFP is a systems scale research facility in the south-west of England for investigation of the sustainability of lowland ruminant production systems (Orr et al., 2016). South-west England has a relatively wet climate where the greatest rainfall is in winter and the driest times are between April to July (Jones et al., 2013). August tends to show an increase in rainfall over July and starts the inexorable rise in rainfall into autumn and early winter. More recently, the number of flood events has increased (Stevens et al., 2016), mostly in the autumn and winter months; all as a likely consequence of increased surface water runoff (Palmer and Smith, 2013).

For this study, the NWFP's 15-minute water flux data were up-scaled to hourly, 6-hourly and daily data and the SPACSYS model was adapted to provide corresponding downscaled simulations at 15-minute, hourly and 6-hourly resolutions (in addition to its usual daily output). This provided four measured water flux datasets and four simulated water flux datasets over a study period of 34 months (April 2013 to February 2016). Simulations were generated using the same field management practices and parameter configurations. These rich water flux datasets enabled investigation of the effects of temporal scale on model performance not only in terms of extreme water runoff, which is the study focus and provides its novelty, but also in terms of general trends.

3.3 Materials and Methods

3.3.1 Model description

The SPACSYS model includes a plant growth and development component, an N cycling component, a C cycling component, a P cycling component, plus a soil water component that includes representation of water flow to field drains as well as downwards through the soil layers, together with a heat transfer component. The equations to quantify such different processes have been described elsewhere (Liu et al., 2013; Wu et al., 2019; Wu et al., 2007; Wu et al., 2015). Here, only the processes influencing directly the soil water component are presented.

For SPACSYS, the Richards' equation for water potential and Fourier's equation for temperature are used to simulate water and heat fluxes, which are inherited from the SOIL model (Jansson, 1998). If the water content in a layer rises above a specified value a proportion is held in macropores such that rapid downward water movement takes place due to gravitational forces alone. Water flow from the soil profile to a drainage pipe occurs when

the ground water table is above the bottom level of the pipe. The Hooghoudt drainage flow equation (Hooghoudt, 1940) with modification is adopted for the subsurface drainage flow. A more detailed description on the SPACSYS model is given in Section 1.8.

The main processes concerning plant growth in SPACSYS are plant development, assimilation, respiration, root growth and development, water uptake, nutrient uptake, biological N fixation for legume plants and partitioning of photosynthate and nutrients from uptake estimated with various mechanisms implemented in the model. N cycling coupled with C cycling covers the transformation processes for organic matter and inorganic N. The main processes and transformations causing size changes to mineral N pools are mineralization, nitrification, denitrification including N gaseous emission and plant N uptake. P cycling is linked to other components such as the plant component, heat transformation and the water cycle. Organic P is subdivided into certain sub-pools with different forms which are connected with transformation rates.

3.3.2 The North Wyke Farm Platform

The study site is located in south-west England, at the NWFP, Rothamsted Research, Okehampton, Devon (50°46’10’’N, 30°54’05’’W). For the period 1985-2015, the mean annual temperature in North Wyke ranges between 6.8 and 13.4 °C, the mean annual rainfall is 1033 mm and the climate is classed as cool temperate. The platform is a 63 ha systems-based experimental facility divided into 15 hydrologically isolated sub-catchments across three 21 ha farmlets with five sub-catchments in each. At the time of this study, all three farmlets were used solely for grazing livestock research (sheep and cattle) where each farmlet was operating under different sward management strategies: no re-seeding (permanent pasture); re-seeded monoculture; and re-seeded legume mix. The platform monitors routinely water runoff and

water chemistry in each of the 15 sub-catchments, together with other primary data collections (e.g. greenhouse gas emissions, livestock performance) so that each farming system can be described, contrasted and compared according to its level of sustainability (Orr et al., 2016). A more detailed description on the NWFP is provided in Section 1.7.

3.3.3 Model configuration

For this study, we focused on water fluxes for one sub-catchment in the permanent pasture farmlet called 'Golden Rove'; a single field that has been under permanent pasture since the outset of the platform in 2010 (Figure 3-1). The soil class for this field is primarily Halstow, which comprises a slightly stony clay loam topsoil (approximately 36% clay) that overlies a mottled stony clay (approximately 60% clay), derived from underlying Carboniferous Culm rocks (Harrod and Hogan, 2008). The study field also has a smaller, but not insignificant area of Denbigh-Cherubeer soil class. In the simulations, the soil type was ignored.

To mimic the grazing system, daily grass intake and excretion of sheep and cattle in the field was estimated before running the simulations (Carswell et al., 2019; Wu et al., 2016). Soil physical and chemical properties of the field were adopted from a previous study of the same field (Wu et al., 2016). The temporal frequency for the measured water fluxes (l s^{-1}) from a NWFP water flume has been 15-minutes since the outset of the platform's setup in October 2012. However, meteorological measurements at the same 15-minute resolution were only initiated from 30 April 2013. Thus, to ensure consistency in the frequency of the driving variables and the water flux as an output variable, simulations also started from 30 April 2013. An end-date of 15 February 2016 was chosen to give an interrupted data collection time period of 34 months. A longer time period would entail having significant periods of missing data due to instrument failure (i.e. there were no measurements on water flux between 15

February and 24 October 2016). A previous scale-focused study, analysing the measured 15-minute water fluxes together with aggregations at 30-minute, hourly, 3-hourly, 6-hourly, 12-hourly and daily resolutions, indicated that 15-minute, hourly, 6-hourly and daily resolutions is sufficient to communicate all key outcomes adequately (Curceac et al., 2020). Thus, the same four temporal resolutions were adopted for the model simulations of this study.

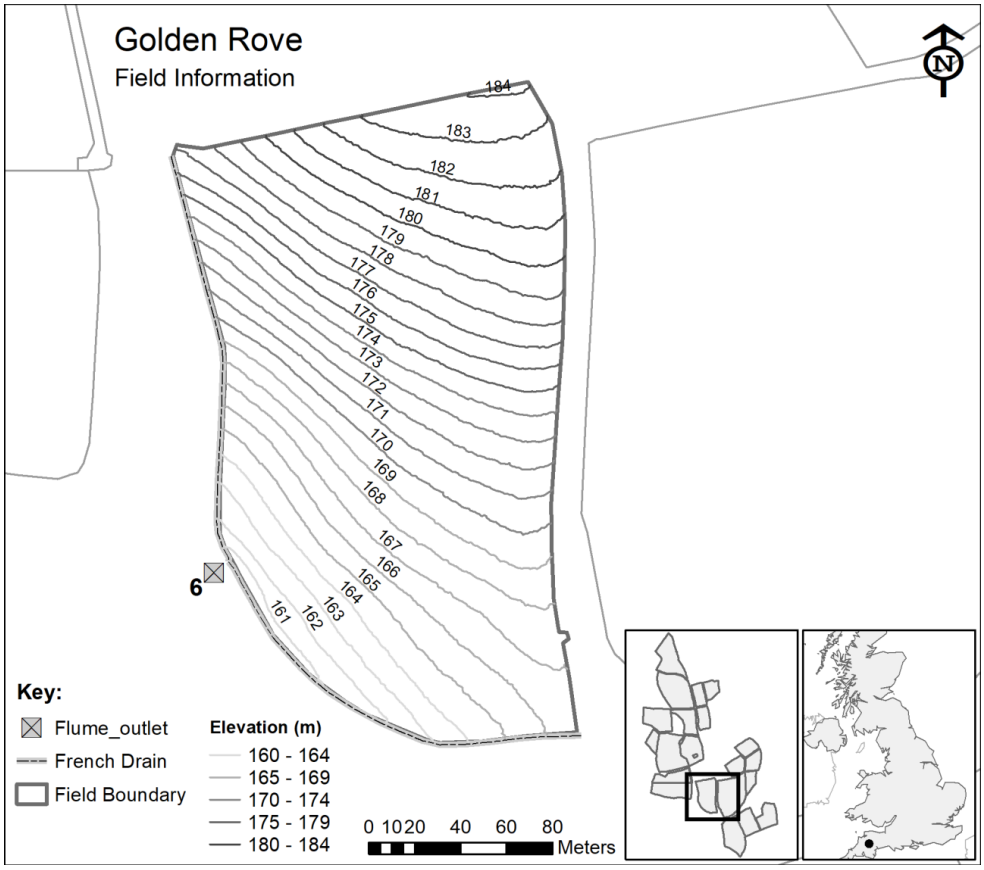


Figure 3-1: Details of the NWFP sub-catchment selected for this study (sub-catchment number 6 of 15, consisting of a single field called Golden Rove). The Rain gauge is approximately centrally located as given in Section 1.7, Figure 1-2.

3.3.4 Statistical analysis

Two distinct sets of statistical analyses were conducted with respect to model performance and data resolution: (a) model performance for each of the four data temporal resolutions,

separately (i.e. 15-minute measured versus 15-minute simulated; hourly measured versus hourly simulated; 6-hourly measured versus 6-hourly simulated; daily measured versus daily simulated); and (b) model performance conducted at the daily temporal resolution only, where 15-minute, hourly and 6-hourly simulations were each aggregated to the daily scale. The latter analyses provide valuable insights into the worth of using fine temporal resolution data to increase the accuracy of daily simulations, especially with respect to the accurate identification of extremes.

3.3.4.1 Model performance graphics

Model performance graphics consist of: (i) time-series plots for measured and simulated datasets; (ii) density plots for measured and simulated datasets; (iii) scatterplots of measured and simulated datasets together with the ideal 1:1 line, a linear regression fit, and a non-linear locally weighted scatterplot smoother (i.e., a Loess smoother fit; Cleveland, 1979); and (iv) time-series plots for the errors (i.e. measured minus simulated data). The estimated intercept and slope parameters from the linear regression fits should equal zero and one for perfect model simulations, respectively. Results (*p*-values) from a linear hypothesis test are reported comparing this ideal model with the estimated model using a finite sample *F* test (see Fox, 2016). The non-linear regression provides added insight into where the simulated values tend to over- or under-predict (e.g., at measured low or high values, respectively).

3.3.4.2 Model performance indices

To further assess the accuracy of the simulations, a set of accuracy statistics was calculated. More specifically, the mean absolute error (MAE), the normalized root mean square error (NRMSE), the percentage bias (PBIAS), the Nash-Sutcliffe efficiency (NSE), the index of

agreement (d) and the Kling-Gupta efficiency (KGE) were computed using the following equations:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{z}_i - z_i| \quad 3\text{-}1$$

$$NRMSE = 100\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{z}_i - z_i)^2}}{z_{max} - z_{min}} \quad 3\text{-}2$$

$$PBIAS = 100\frac{\sum_{i=1}^{N}(\hat{z}_i - z_i)}{\sum_{i=1}^{N} z_i} \quad 3\text{-}3$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}(\hat{z}_i - z_i)^2}{\sum_{i=1}^{N}(z_i - \bar{z}_i)^2} \quad 3\text{-}4$$

$$d = 1 - \frac{\sum_{i=1}^{N}(\hat{z}_i - z_i)^2}{\sum_{i=1}^{N}(|\hat{z}_i - \bar{z}_i| + |z_i - \bar{z}_i|)^2} \quad 3\text{-}5$$

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{\hat{z}}}{\sigma_z} - 1\right)^2 + \left(\frac{\bar{\hat{z}}}{\bar{z}} - 1\right)^2} \quad 3\text{-}6$$

where $\hat{z}_i$ are the simulated values, z_i are the observations, $\bar{z}_i$ is the mean of the measured values, r is the Pearson product-moment correlation coefficient (between measured and simulated) and σ is the standard deviation. The optimal value of the error (i.e. model residual) indices (MAE, NRMSE and PBIAS) is zero such that the smaller the value, the more accurate the model simulation. NSE takes values from $-\infty$ to 1, where unity corresponds to a perfect match between the measured and simulated values, zero indicates that the model simulations are as accurate as the mean of the measured values and a negative value indicates that the mean of the observations is a better predictor than the model. The index of

agreement d is defined in the range 0 to 1, where again unity shows perfect model performance and zero, no agreement at all. KGE incorporates r, the ratio between the means of observations and simulations and the variability ratio. It has the same range as the NSE.

3.3.4.3 Simulation accuracy of measured peaks

To investigate model accuracy in simulating water flux peaks, a threshold at the 99th percentile of each measured water flux dataset was used to identify peak flows. Model simulations were then assessed to determine if they could similarly exceed this threshold coinciding with an measured exceedance. Incidences of correct peak flow simulations, false negatives (simulation does not exceed threshold when measured flow does), false positives (simulation exceeds threshold when measured flow does not) and corresponding kappa values are reported. The kappa statistic provides a measure of agreement beyond the level of agreement expected by chance alone. General guidelines for kappa values are as follows: less than 0.2 slight agreement, 0.2 to 0.4 fair agreement, 0.4 to 0.6 moderate agreement, 0.6 to 0.8 substantial agreement, greater than 0.8 almost perfect agreement, and equal to 1 perfect agreement.

3.4 Results

3.4.1 Model performance for each of the four data temporal resolutions, separately

Comparisons between the measured and simulated water flux rates at different temporal resolutions are shown in Figure 3-2. Visually, it appears that simulations of daily and 6-hourly water fluxes tend to under-predict the measured data, often missing high peaks, while simulations of 15-minute and hourly data possibly tend to over-predict. Figure 3-3 shows the link of this behaviour to the seasonality of the rainfall. However, the scatterplots of the measured and simulated data, together with the ideal 1:1 line, a linear regression fit, and a

Loess smoother fit (Harrell, 2001) (Figure 3-4) present a more complete picture. Simulations for all four temporal resolutions clearly tend to over-predict, with the level of over-prediction increasing as the resolution increases. Over-prediction is shown with each linear regression fit lying below the 1:1 line; and increasingly so, as the resolution increases. All linear regression fits were found to be significantly different to the 1:1 line, each with *F*-test *p*-values < 0.0001 (Table 3-1).

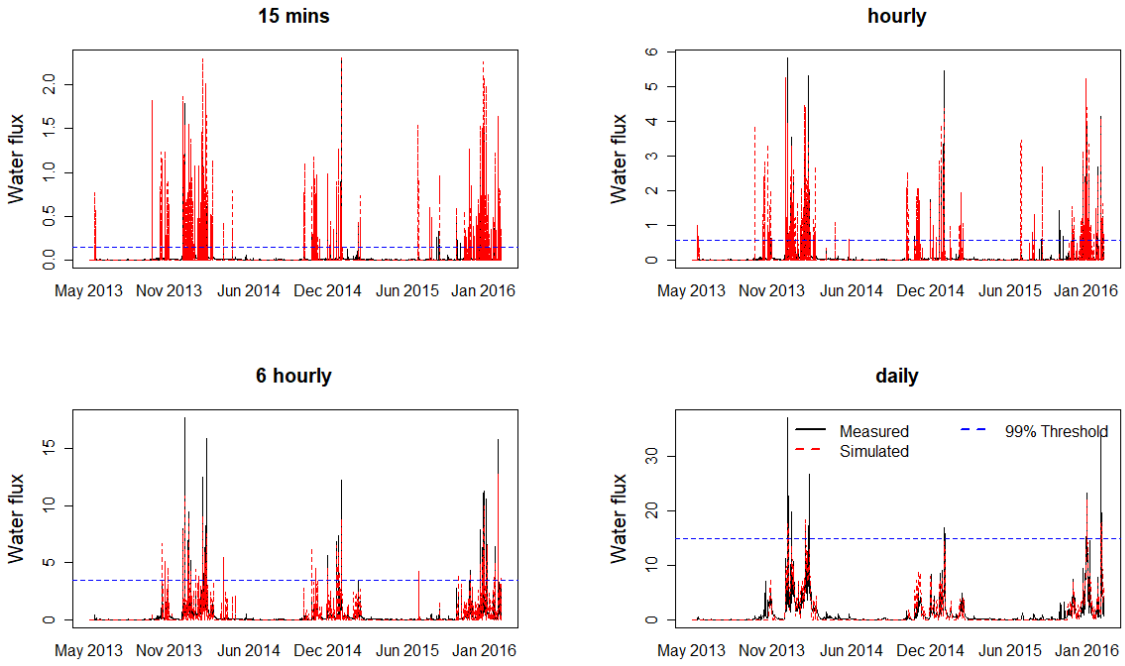


Figure 3-2: Time-series plots for measured and simulated water flux data (not aggregated) for 15-minute, hourly, 6-hourly and daily data (in units of mm 15min^{-1}, mm h^{-1}, mm 6h^{-1} and mm d^{-1}, respectively). All plots are shown with a threshold at the 99th percentile of measured data (at 0.138 mm 15min^{-1}, 0.553 mm h^{-1}, 3.45 mm 6h^{-1} and 14.9 mm d^{-1}, respectively).

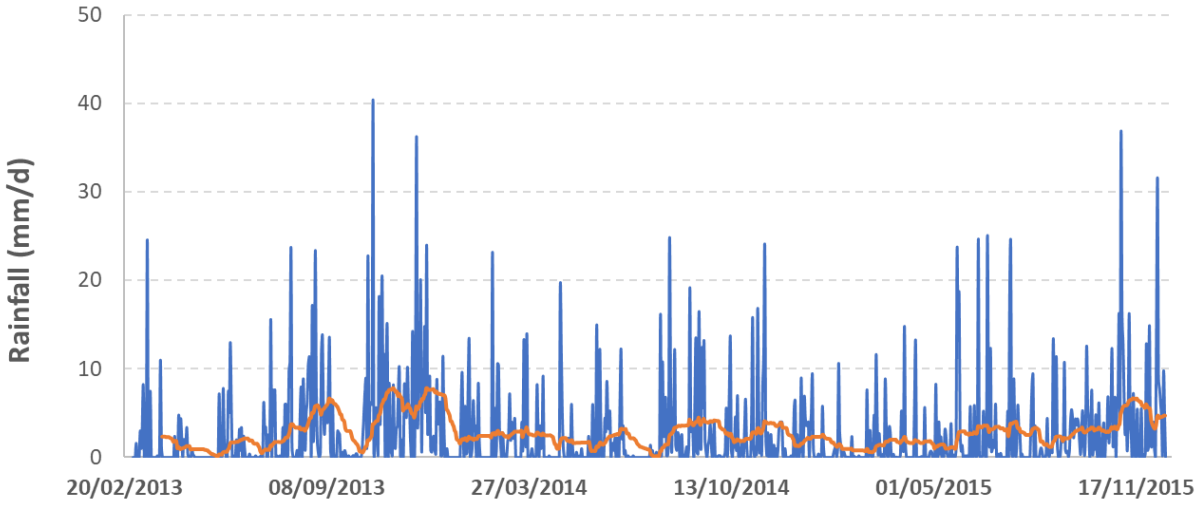


Figure 3-3: Time series of measured daily rainfall (mm d^{-1}) with a monthly moving average.

For all four temporal resolutions, the tendency to over-predict decreases at the largest measured water fluxes, as shown by the concave behaviour of the loess smoother fit (Figure 3-4), with daily simulations tending to under-predict at very large fluxes, thus, missing extreme events that may cause flooding and associated high nutrient and sediment losses. Clearly, 'smoothing bias' increases as temporal resolution decreases. The 15-minute simulations maintain the variation shown in the measured data (i.e. observations range from 0 to 2.306 mm $15min^{-1}$ while simulations range from 0 to 2.310 mm $15min^{-1}$), while the daily simulations do not (i.e. observations range from 0 to 36.97 mm d^{-1} while simulations only range from 0 to 22.20 mm d^{-1}). As each 'simulation-to-observation' comparison is on a different scale, it is not useful to present further model fit diagnostics, such as error and agreement indices.

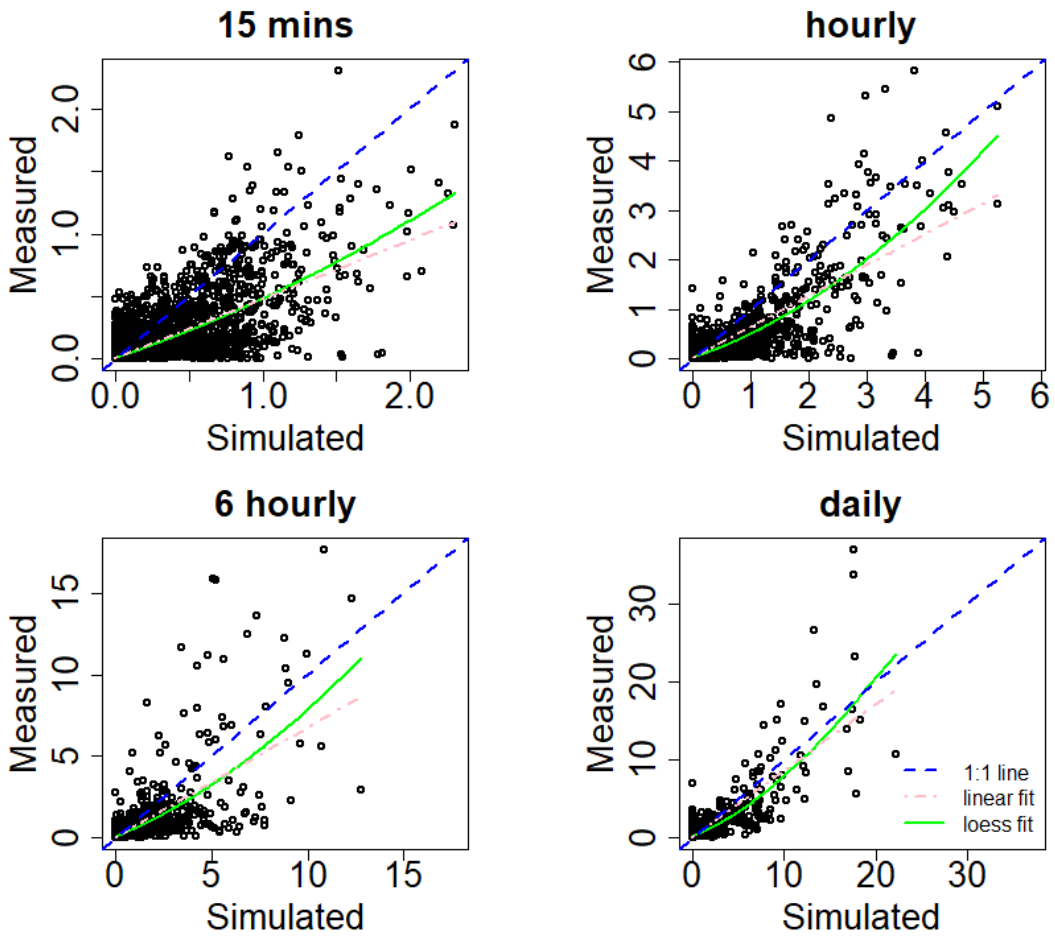


Figure 3-4: Scatterplots of the measured and simulated data (not aggregated) for 15-minute, hourly, 6-hourly and daily data. Scatterplots are shown with the 1:1 line, a linear regression fit and a loess smoother fit. Units are in mm 15min^{-1}, mm h^{-1}, mm 6h^{-1} and mm d^{-1}, respectively.

Table 3-1: Regression coefficients and R^2 values for measured and simulated flow at 15-minute, hourly, 6-hourly and daily scale.

	Intercept estimate	Slope estimate	R^2
15mins	0.006	0.473	0.554
hourly	0.017	0.628	0.677
6 hourly	0.042	0.675	0.504
daily	0.070	0.856	0.685

3.4.2 Model performance for simulations aggregated to the daily scale

Comparisons between the measured and simulated water flux rates aggregated to the same daily scale are shown in Figure 3-5. Since the data is expressed in mm per time unit, the aggregation was done by summing up (96 values at 15-minute, 24 values at hourly and 4 values at 6-hourly scale) instead of calculating averages if the data was expressed in l s^{-1}. There are clear instances of both over- and under-prediction for all four daily outputs. The scatterplots (Figure 3-6) of daily measured and daily simulated data from different aggregations, together with the 1:1 line, a linear fit, and a loess smoother fit, again provide a clear visualisation of the relations in the time-series plots. Simulations for all three aggregations to daily (15-minute, hourly and 6-hourly) again tend to over-predict (as their linear fits lie below the 1:1 line), but this over-prediction is broadly similar across the four datasets, and not as great as that found with the unaggregated data, above. The 6-hourly aggregations appear to be the least accurate. Again, all linear regression fits were found to be significantly different to the 1:1 line, each with *F*-test *p*-values < 0.0001.

In this instance, 'smoothing bias' increases as aggregation resolution decreases, where simulations for 15-minute and hourly aggregations both increase the variation shown in the measured daily data (i.e. 0 to 36.97 mm d^{-1}); with 15-minute daily aggregations ranging from 0 to 38.94 mm d^{-1} and hourly daily aggregations ranging from 0 to 39.64 mm d^{-1}. Conversely, the 6-hourly daily aggregations and the daily simulations reduce variation with the 6-hourly daily aggregations ranging from 0 to 31.70 mm d^{-1} and the (unaggregated) daily simulations ranging from 0 to 22.20 mm d^{-1}. In summary, daily simulations based on component 15-minute and hourly aggregations have the potential to identify peak water fluxes (and, thus, flood events) and predict their magnitudes more accurately, relative to 6-hourly

aggregations and (unaggregated) daily simulations.

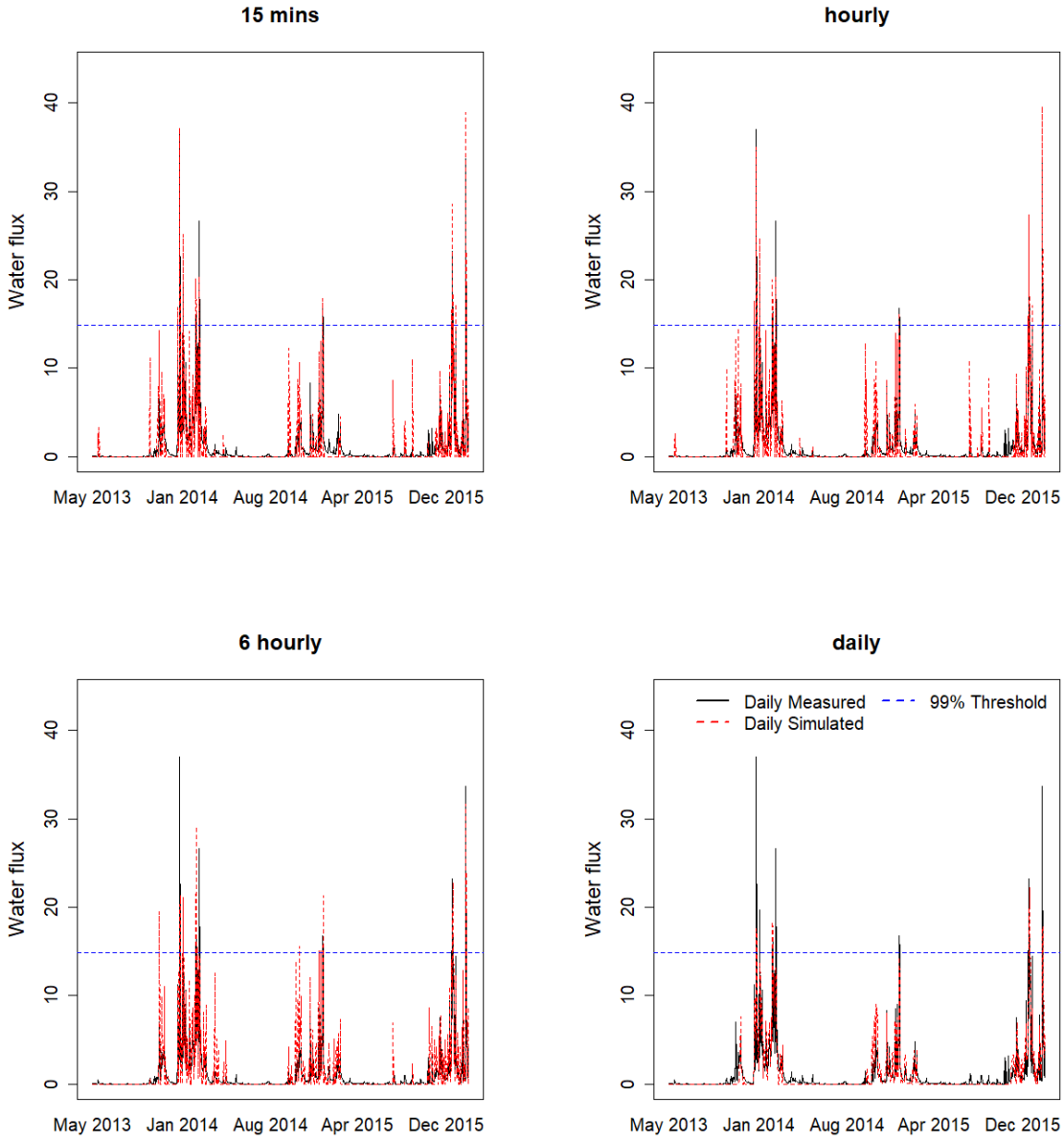


Figure 3-5: Time-series plots for daily measured and daily simulated water flux data (with the first three plots having data aggregated from: 15 minutes to daily; hourly to daily; 6 hourly to daily). All units in mm d^{-1}. All plots are shown with a threshold at the 99th percentile of measured data (14.90 mm d^{-1}).

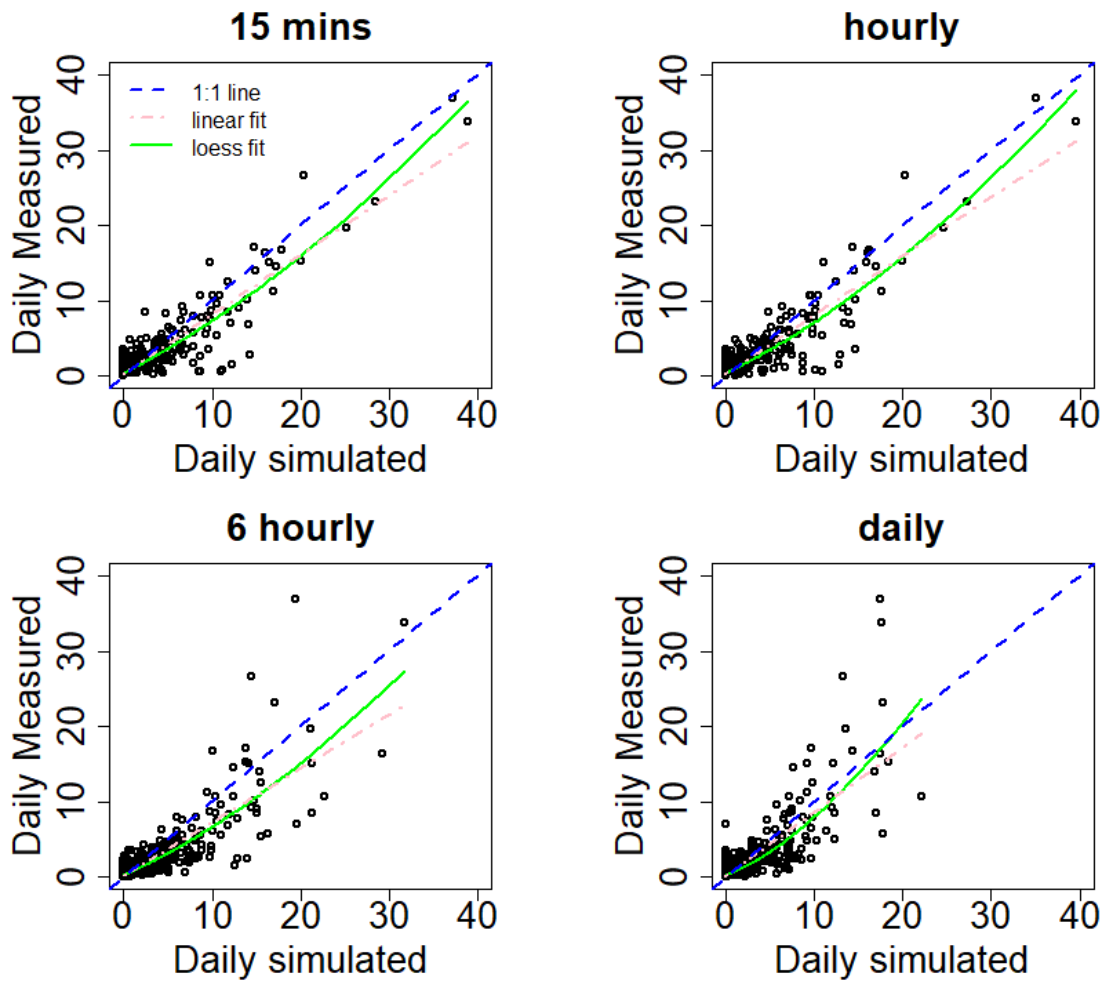


Figure 3-6: Scatterplots of the daily measured and daily simulated data (with the first three plots having data aggregated from: 15 minutes to daily; hourly to daily; 6 hourly to daily). Scatterplots are shown with the ideal 1:1 line, a linear regression fit and a loess smoother fit. All units in mm d^{-1}.

Further clarity on bias is provided in the density plots for the measured and simulated data (Figure 3-7and Figure 3-8). Here, daily simulations based on 15-minute and hourly aggregations have a lower density at small daily water fluxes than that found with the measured data, while the 6-hourly aggregations and (unaggregated) daily simulations have a higher density at small daily water fluxes. This is combined with a longer tail in the density curve for the 15-minute and hourly aggregations, as each can simulate large daily water

fluxes, while the 6-hourly aggregation and (unaggregated) daily simulations do not have this property.

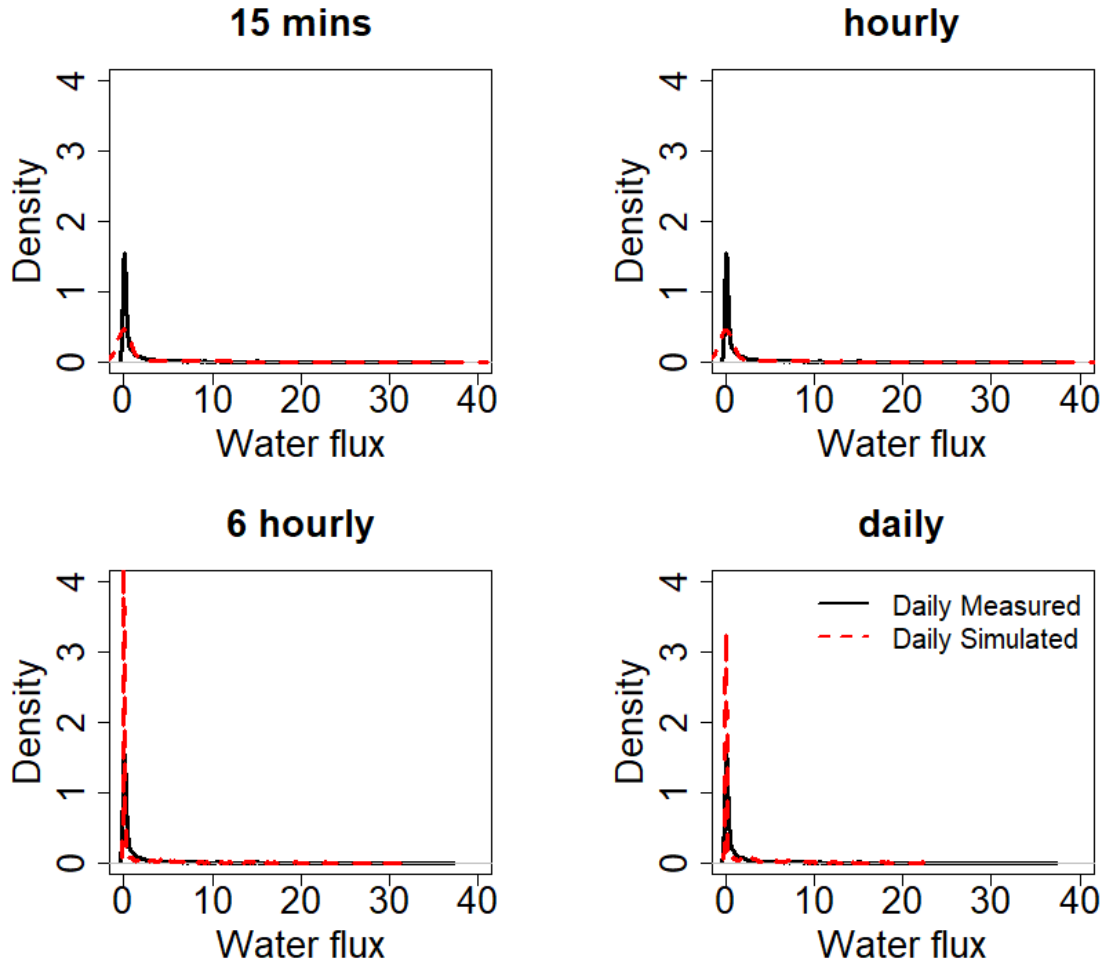


Figure 3-7: Probability density plots for daily measured and daily simulated data (with the first three plots having data aggregated from: 15 minutes to daily; hourly to daily; 6 hourly to daily). All units in mm d^{-1}.

/

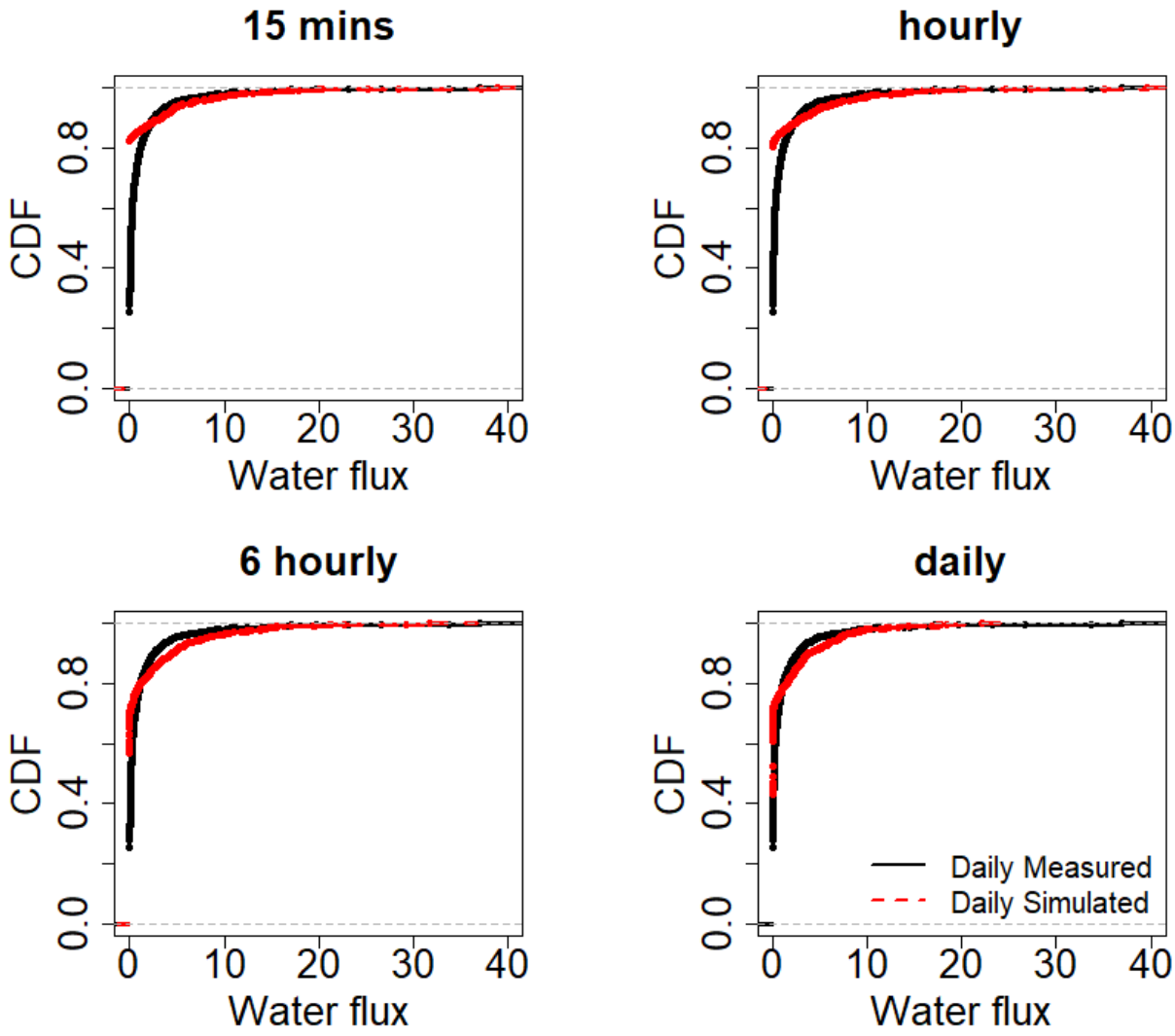


Figure 3-8: Cumulative density plots for daily measured and daily simulated data (with the first three plots having data aggregated from: 15 minutes to daily; hourly to daily; 6 hourly to daily). All units in mm d^{-1}.

The cumulative flow (Figure 3-9) shows that during the first flows around November 2013, the flow is over-predicted by the 15-minutes and hourly simulations (aggregated to daily) while the daily simulations under-predict. This behaviour is then reversed as the daily simulations over-predict and the 15-minutes and hourly simulations under-predict until the end of the considered period. The 6-hourly simulations are considerably more positively biased.

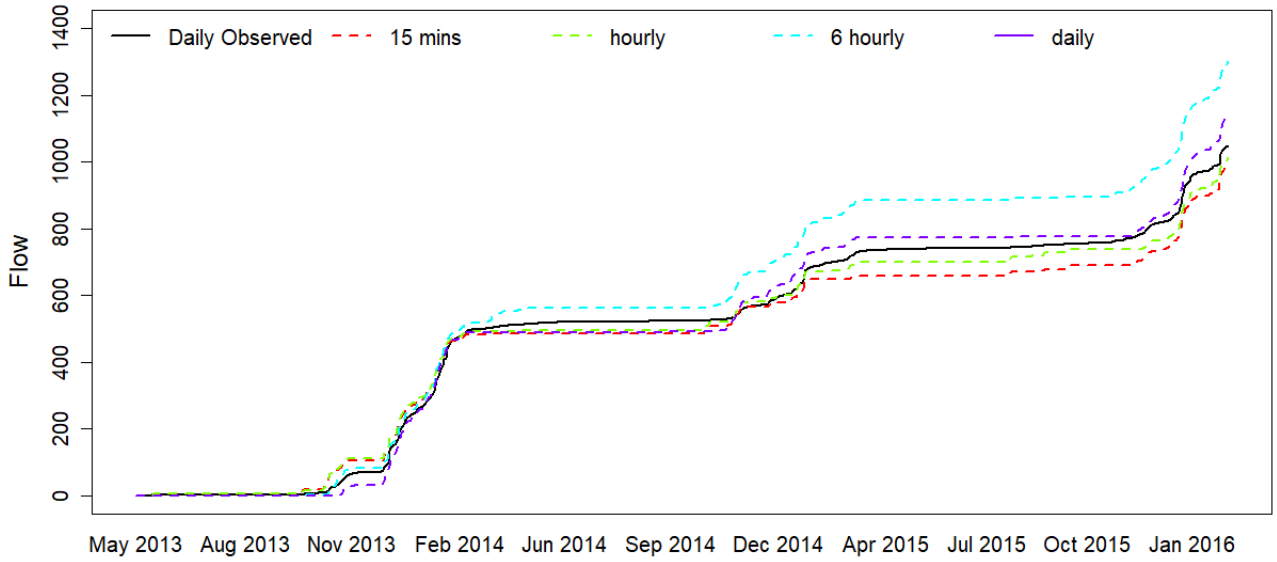


Figure 3-9: Cumulative daily measured flow and 15-minutes, hourly, 6-hourly and daily simulated flow aggregated to daily for the whole considered period.

The error indices (MAE, NRMSE and PBIAS) are reported for each daily aggregation in Figure 3-10, where the 15-minute and hourly aggregations clearly perform more accurately than the 6-hourly aggregation and (unaggregated) daily simulations. Errors (i.e. residuals) are also reported over the study time period in Figure 3-11, where errors tend to be larger for the daily simulations based on the 6-hourly aggregation and the (unaggregated) daily simulations. Interestingly, the 6-hourly aggregation consistently is the least accurate, including being less accurate than the (unaggregated) daily simulations.

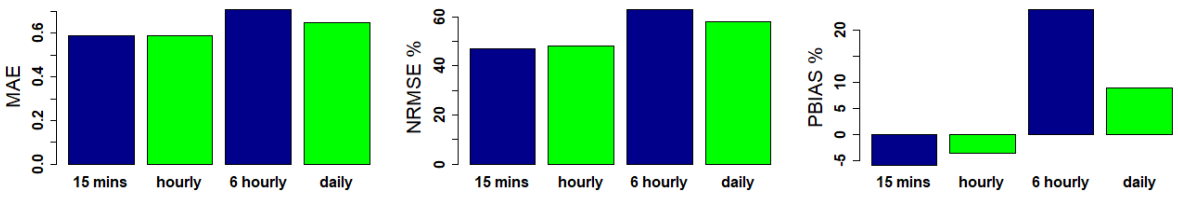


Figure 3-10: Error (MAE, NRMSE, PBIAS) indices with respect to daily measured and daily simulated data (with 15-minute, hourly, 6-hourly data aggregated to daily).

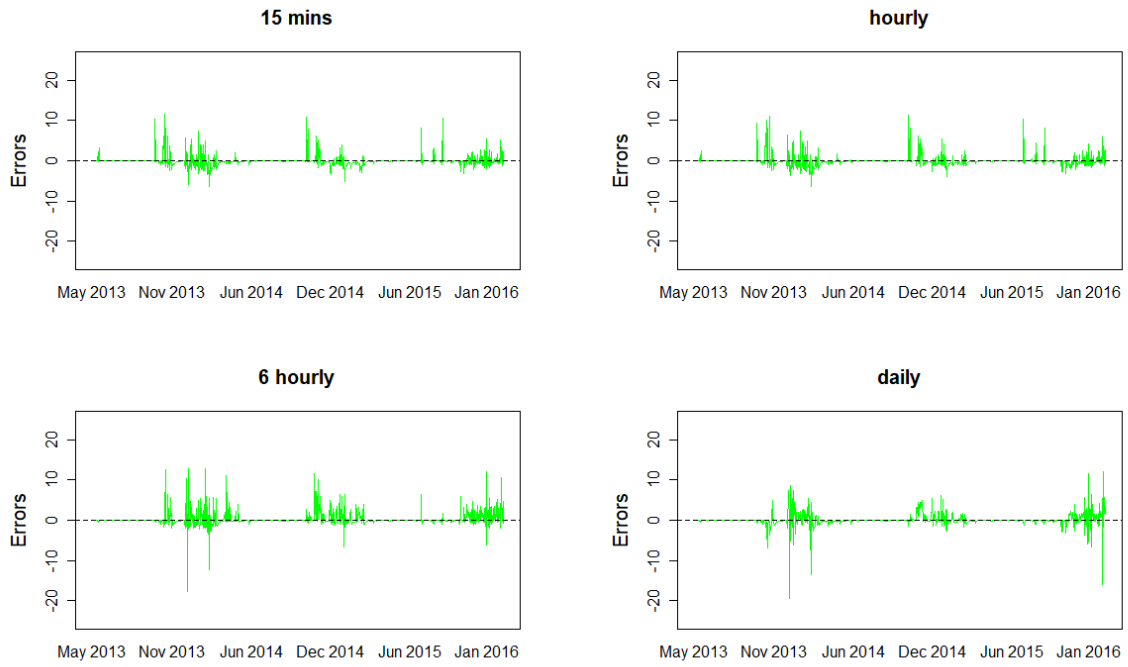


Figure 3-11: Time-series of errors (simulated minus measured data) aggregated to the daily temporal resolution. All units in mm d^{-1}. Positive errors represent over-prediction by the model.

Agreement indices (NSE, d and KGE) are reported for each daily aggregation in Figure 3-12, where again the 15-minute and hourly aggregations perform more accurately than the 6-hourly aggregation and (unaggregated) daily simulations (although daily simulations perform relatively well according to the KGE index). From the given accuracy diagnostics, it is not immediately apparent whether daily simulations based on 15-minute or hourly aggregations are the most accurate, and as such, both appear to increase the accuracy relative to SPACSYS' daily simulations. Again, the 6-hourly aggregation is the least accurate.

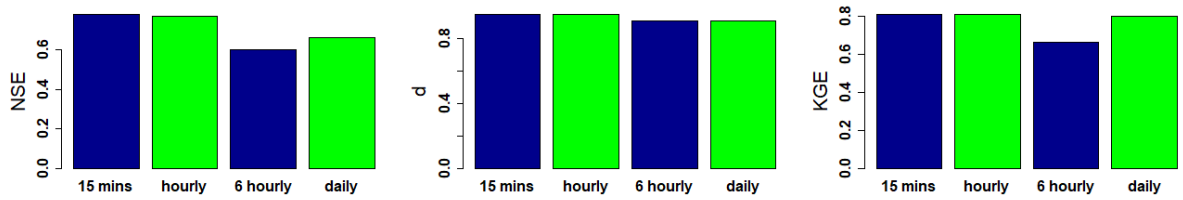


Figure 3-12: Agreement (NSE, *d*, KGE) indices with respect to daily measured and daily simulated data (with 15-minute, hourly, 6-hourly simulations aggregated to daily).

3.4.3 Simulation of measured peaks

To investigate the ability of the model to simulate and identify water flux peaks, the 99th percentile of each measured water flux dataset was used as a threshold to identify peak flows, as highlighted in Figure 3-2 and Figure 3-5 (the dashed blue line). Incidences of correct peak flow simulations, false negatives, false positives and the resultant kappa values are given in Table 3-2. It appears that hourly simulations provide the largest correct classification rate (kappa = 0.553) for the unaggregated approach, but with only moderate success in identifying measured peak flows (as a promising 92% identification rate is tempered by a poor mis-identification rate). Conversely, aggregating to the daily scale using either 15-minute or hourly simulations was able to provide much greater agreement in identifying measured peak flows at the daily scale, with each identifying 9 out of 11 peak flow events correctly, coupled with only 2 false positives (kappa = 0.816 in both cases). This level of agreement was far greater than that found through directly simulating the daily data, which provided only moderate success in identifying measured peak flows (kappa = 0.495). Again, the 6-hourly aggregation is the least accurate with a relatively high number of false positives (simulated flow exceeds the threshold when measured flow does not).

Table 3-2: Accuracy at peak water fluxes according to simulation resolution. Peaks taken at 99th percentile of measured data (see the dashed blue line in Figure 3-2 and Figure 3-5).

Simulation resolution	Sample size	Measured Peaks	Correctly Simulated	False Negative	False Positive	kappa
Unaggregated						
15-minute	97920	980	759 (77)*	221	1224	0.506
hourly	24480	245	225 (92)	20	335	0.553
6-hourly	4080	41	32 (78)	9	52	0.503
daily	1020	11	5 (45)	6	4	0.495
Aggregated to daily						
15-minute	1020	11	9 (82)	2	2	0.816
hourly	1020	11	9 (82)	2	2	0.816
6-hourly	1020	11	6 (55)	5	9	0.455
daily	1020	11	5 (45)	6	4	0.495

* Value in brackets shows a percentage of correctly simulated peaks to measured peaks.

3.5 Discussion

3.5.1 Model performance

3.5.1.1 Unaggregated data

The statistical analyses for model performance suggested that the SPACSYS model simulates the general trend of water fluxes at the four different temporal resolutions reasonably well (Figure 3-2 and Figure 3-4). All simulations tended to over-predict water flux, however, and only simulations at the finest resolutions maintained the variation in the measured data. The accuracy of water flux peak simulations varied among the four resolutions (Table 3-2). Almost

92% of the measured peaks over the simulated period were modelled correctly at an hourly resolution, the resolution with the smallest misclassification rate. However, this was tempered by a high rate of predicting peaks that did not exist. A previous statistical analysis of peak flows at different scales from a different NWFP sub-catchment (similar in size to the one used here), modelled and simulated by a Generalized Pareto distribution, also showed the greatest agreement at the hourly resolution (Curceac et al., 2020).

3.5.1.2 Aggregated to Daily

When simulations at a finer temporal resolution were aggregated to a daily rate, the simulation results using both the 15-minute and hourly aggregations showed the greatest accuracy broadly equally, both in the prediction of general trends (Figure 3-5 to Figure 3-12) and the identification of peak flows (Table 3-2). This demonstrates clearly that the daily simulation of water fluxes with the SPACSYS model, informed by finer temporal resolution data, can increase simulation accuracy. This result is an important advance relative to previous SPACSYS studies, which only used a daily time-step, and which similarly used sub-catchments of the NWFP as the study site (Liu et al., 2018).

3.5.2 Results in context and their generalisation

Results are consistent with other studies that similarly showed that differences in the (unaggregated or aggregated) time-step have the greatest impact on runoff simulation accuracy relative to other factors, some of which, also investigated changes in spatial resolution (i.e. aggregating over different spatial units) (Choi et al., 2018; Huang et al., 2019; Jeong et al., 2010; Kavetski et al., 2011; Merz et al., 2009). A possible explanation for this behaviour could be that the infiltration-runoff partitioning is more accurately described at higher temporal and spatial resolutions (Regenass et al., 2021). Thus, the value of using

aggregated fine temporal resolution simulations to increase the accuracy of daily simulations can be said to hold generally for other process-based models provided the hydrological process is described appropriately. However, it does not follow that daily simulation accuracy will continue to increase as the temporal resolution of the aggregated data becomes finer. This study found aggregating hourly simulations to daily to be just as accurate as aggregating 15-minute simulations to daily, while aggregating 6-hourly simulations to daily performed less well than the usual daily simulations.

The temporal resolution for process-based models should be chosen carefully to balance between capturing all important processes, the study objectives and data availability. For our study, with flooding as context, the identification of water flux extremes in a grassland field (or small sub-catchment) with a heavy clay soil, is viewed as *the* important process, more so than capturing broad trends in water flux. It is well-known that running a model at a finer resolution, then aggregating, will increase the prediction accuracy in a broad sense (see above). What has received less attention in the literature is the effects of temporal resolution on a model's ability to capture extremes (e.g. see Schaller et al., 2020, in the context of streamflow). In this respect, our study has found daily peak flows to be more accurately identified using aggregations of simulations at finer resolutions, than using coarse daily simulations directly. Of course, measurement at a finer resolution comes at a cost and this needs to be balanced with associated improvements in model accuracy. In this instance, this interplay is simple to resolve since aggregating to the daily scale using both 15-minute and hourly simulations were equally as accurate, meaning measuring at an hourly interval would be sufficient for the case study site.

The appropriate temporal resolution to simulate water fluxes using a hydrological model depends on hydro-climatological and geophysical characteristics, and the scale of the process. It has been suggested that an appropriate temporal resolution could be between 12 hours for middle-sized upstream areas and 48 hours for a complete river basin (Booij and Tran, 2005). As the size of the field for this study is < 4 ha, the indicated hourly resolution appears reasonable. Observed and projected changes in the UK's climate suggest an increase in heavy rain events and wetter winters (Committee on Climate Change, 2017), where some UK regions will be more affected than others. This will inevitably change agricultural management practice and land use across the UK. Taking as an example the grazed pasture of this study, introducing a deep-rooting grass suited to its heavy clay soils (Macleod et al., 2013) and/or the mechanical loosening of topsoil (Newell-Price et al., 2011) would probably reduce overland runoff, whereas conversion to an arable crop (e.g. wheat) would provide its own set of water runoff influences, where both total runoff and peak flow would be expected to increase (Gerla, 2007; Ahiablame et al., 2019). Such changes would alter the characteristics of the water fluxes generated, as the field's soil properties will change, meaning the determination of an appropriate resolution to simulate water fluxes may also change from the hourly resolution suggested here. This is analogous to other hydrological studies where, for example, different overflow designs in roof drainage structures have markedly variable responses to rainfall intensity increases (Verstraten et al., 2019).

3.5.3 Inputs that impact hydrological model performance

Key model input variables such as precipitation can determine the impacts of simulation time-steps on the performance of hydrological models; for example, the duration and temporal variability of a precipitation event in relation to the rainfall–runoff lag time (Ficchì et al.,

2016). A multiple-day precipitation event is the main cause of continuous runoff events and related peaks. For example, for this study, there was almost an unbroken measured precipitation period from 14 December 2013 to 5 March 2014, which brought a total of 541 mm of precipitation, 78% of which was measured as surface runoff (i.e., measured water flux). Study simulations showed 70, 70, 81 and 85% as water fluxes over the period at the 15-minute, hourly, 6-hourly and daily resolutions, respectively. Previous studies showed that wetter soils had less capacity to store water, resulting in greater runoff volumes (Huang et al., 2017; Kibet et al., 2014; Zehe et al., 2010). Both observations and simulations in this study confirmed this finding. Conversely, for a single day event, a measured 92% of 40.2 mm daily precipitation was discharged on 23 December 2013. The simulations generated 99, 99, 90 and 44% water losses at the 15-minute, hourly, 6-hourly and daily resolutions, respectively. Thus, almost all of the precipitation contributed to the water loss on this day, where only the daily-scale simulation did not capture this. However, although heavy rainfall is necessary to generate water fluxes, it is not a sufficient condition for a higher surface runoff rate to occur (Ledingham et al., 2019). For example, there was about 25 mm precipitation on 14 May 2013 and on 13 August 2015, but both the simulations and the observations (at all four resolutions) did not show apparent water fluxes. Further, a daily precipitation of 4 mm on 27 February 2015 generated a measured 120% water loss, together with simulated values of 48, 148, 125 and 75% water loss at the 15-minute, hourly, 6-hourly and daily resolutions, respectively.

The generation of water fluxes not only depends on the intensity of precipitation, but also surface coverage, topography and soil physical properties of the field. In hydrology, lag time, defined as the time difference between the peak runoff and mass centre of rainfall excess (Hall, 1984), is usually used to determine a runoff rate. Although the SPACSYS model does not

use this parameter to estimate the surface runoff rate, it uses the Richards' equation to calculate soil water redistribution where soil hydraulic conductivity, saturated water content and plant uptake play critical roles in water movement and consequently runoff. A trial study on the spatial variation of soil hydraulic conductivity (unpublished data) in a NWFP field, nearby to the study field, highlighted clear within-field variation, partially because of compaction caused by grazed animal movement. However, the soil physical properties used in the simulations were estimated based on soil texture, and at the field-scale only. The error and uncertainty introduced by this approach are likely to be transferred to the errors in simulating infiltration and surface runoff rates. To improve model simulation accuracy, soil physical properties as core information should be provided wherever possible.

3.5.4 Further considerations of scale

The processes controlling water fluxes operate across a range of spatial and temporal scales, and the time-series that are recorded represent an aggregation of these effects. For example, effects of evapotranspiration will dominate at annual scales whereas more local impacts of precipitation will manifest at finer scales (Rust et al., 2014). As noted above, the characteristics of the study area, (e.g. size, soil condition and topography) will impact the dominant scales of variation and hence the frequency at which it is most appropriate to model or measure water fluxes. Rust et al. (2014) presented an analysis which aimed to determine whether the process-based model they studied captured the scale-dependent variation measured in catchment runoff. They analysed model residuals using wavelet-based signal processing methods and found that although their model captured broadly the scale-dependent variation in the data, fine scale variation was always under-predicted. Our results

shed light on their observation as we confirm that if the scale of the model prediction is not sufficiently fine then model-damping will result in an underprediction of extreme events.

3.6 Conclusions

For the grassland study site, the adapted process-based model (SPACSYS) could adequately simulate the trends in measured water fluxes and identify their extremes. At a daily time-step, model accuracy increased when simulations were run at finer temporal resolutions, specifically 15-minute and hourly, and then aggregated to daily (a coarse output resolution commonly used in field-scale agricultural settings). Aggregating using 6-hourly simulations was less accurate. For the study site, which constitutes a field of a grassland research farm platform (NWFP), simulation of water fluxes at an hourly resolution is likely optimal since use of the 15-minute resolution did not increase prediction accuracy or the ability to identify extremes in flow further. Therefore, for modelling purposes, monitoring frequency could be reduced safely to hourly from the current 15-minute resolution.

Results provide information not only for the NWFP experiment, but also and indirectly, the UK grassland farming regions that its outputs upscale to (Pulley and Collins, 2019). Study results are crucial in relation to meeting the increasing demand for reliable simulation-based runoff forecasts at daily and sub-daily resolutions, where accurate knowledge of peak discharge and stage are essential not only for flood protection, but also to help increase the forecast accuracy of associated emissions such as nutrient or sediment loss, that each use water flux as a component. Further research is called for in specifying the temporal resolution amongst the wide range of field-scale hydrological/agricultural models currently applied. This needs to be coupled with linked changes in climate and land use to increase model forecast accuracy and to optimise data acquisition schemes on farms generally.

Authors contribution statement

Lianhai Wu 45%: Conceptualization, Methodology, Software, Writing – original draft, Writing - review & editing, Supervision, Funding acquisition.

Stelian Curceac 30%: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing.

Peter M. Atkinson 5%: Writing - review & editing, Supervision, Funding acquisition.

Alice Milne 5%: Writing - review & editing, Supervision, Funding acquisition.

Paul Harris 15%: Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition.

Acknowledgements

This research was funded by the BBSRC Institute Strategic Programme grant, "Soils to Nutrition" (BBS/E/C/000I0330, BBS/E/C/000I0320), the BBSRC National Capability grant for the North Wyke Farm Platform (BBS/E/C/000J0100) and a PhD studentship funded by Rothamsted Research and Lancaster University.

References

Abalos, D., Cardenas, L.M., Wu, L., 2016. Climate change and N_2O emissions from South West England grasslands: a modelling approach. Atmos. Environ., 132: 249-257. DOI:https://doi.org/10.1016/j.atmosenv.2016.03.007

Adimassu, Z., Alemu, G., Tamene, L., 2019. Effects of tillage and crop residue management on runoff, soil loss and crop yield in the Humid Highlands of Ethiopia. Agric. Syst., 168: 11-18. DOI:https://doi.org/10.1016/j.agsy.2018.10.007

Ahiablame, L., A. Y. Sheshukov, E. Mosase and J. Hong. (2019). Modelling the impacts of grassland to cropland conversion on river flow regimes in Skunk Creek watershed, Upper Midwest United States, 35 (9), 1454-1465. https://doi.org/10.1002/rra.3512.

Ahuja, L.R., Ma, L., Howell, T.A., 2002. Agricultural System Models in Field Research and Technology Transfer. CRC Press, Boca Raton.

Alaoui, A., Rogger, M., Peth, S., Blöschl, G., 2018. Does soil compaction increase floods? A review. J. Hydrol., 557: 631-642. DOI:https://doi.org/10.1016/j.jhydrol.2017.12.052

Archer, D.R., Fowler, H.J., 2018. Characterising flash flood response to intense rainfall and impacts using historical information and gauged data in Britain. J. Flood Risk Manag., 11(S1): S121-S133. DOI: https://doi.org/10.1111/jfr3.12187

Bingham, I.J., Wu, L., 2011. Simulation of wheat growth using the 3D root architecture model SPACSYS: validation and sensitivity analysis. Eur. J. Agron., 34: 181-189. DOI:https://doi.org/10.1016/j.eja.2011.01.003

Booij, M.J., Tran, H.M., 2005. Appropriate scales for hydro-climatological variables in the Red River basin. In: Wagener, T. et al. (Eds.), Proceedings of symposium S6 held during the Seventh IAHS Scientific Assembly. IAHS publication. IAHS Press, pp. 325-332.

Brown, I., Thompson, D., Bardgett, R., Berry, P., Crute, I., Morison, J., Morecroft, M., Pinnegar, J., Reeder, T., Topp, K., 2016. UK Climate Change Risk Assessment Evidence Report: Chapter 3, Natural Environment and Natural Assets. Report Prepared for the Adaptation Sub-committee of the Committee on Climate Change, London.

Carswell, A.M., Gongadze, K., Misselbrook, T.H., Wu, L., 2019. Impact of transition from

permanent pasture to new swards on the nitrogen use efficiency, nitrogen and carbon budgets of beef and sheep production. Agric. Ecosyst. Environ., 283: 106572. DOI:https://doi.org/10.1016/j.agee.2019.106572

Charlton, M.B., Bailey, A., Arnell, N., 2010. Water for Agriculture – Implications for Future Policy and Practice Report for the Royal Agricultural Society of England.

Choi, Y.S., Shin, M.-J., Kim, K.T., 2018. Preliminary study of computational time steps in a physically based distributed rainfall–runoff model. Water, 10(9): 1269. DOI: https://doi.org/10.3390/w10091269

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc., 74(368): 829-836. DOI:https://10.1080/01621459.1979.10481038

Committee on Climate Change, 2017. UK Climate Change Risk Assessment 2017. Synthesis Report., London.

Curceac, S., Atkinson, P.M., Milne, A., Wu, L., Harris, P., 2020. An evaluation of automated GPD threshold selection methods for hydrological extremes across different scales. J. Hydrol., 585: 124845. DOI:https://doi.org/10.1016/j.jhydrol.2020.124845

Dadson, S.J., Hall, J.W., Murgatroyd, A., Acreman, M., Bates, P., Beven, K., Heathwaite, L., Holden, J., Holman, I.P., Lane, S.N., O'Connell, E., Penning-Rowsell, E., Reynard, N., Sear, D., Thorne, C., Wilby, R., 2017. A restatement of the natural science evidence concerning catchment-based 'natural' flood management in the UK. Proc. R. Soc. A, 473(2199): 20160706. DOI:https://doi.org/10.1098/rspa.2016.0706

Del Grosso, S.J., Ojima, D.S., Parton, W.J., Stehfest, E., Heistemann, M., DeAngelo, B., Rose, S., 2009. Global scale DAYCENT model analysis of greenhouse gas emissions and mitigation strategies for cropped soils. Glob. Planet. Change, 67(1): 44-50. DOI:https://doi.org/10.1016/j.gloplacha.2008.12.006

Department for Environment, Food and Rural Affairs, 2019. Agriculture in the United Kingdom 2018, London.

Environment Agency, 2009. Flooding in England: A National Assessment of Flood Risk, Bristol, UK.

Ficchì, A., Perrin, C., Andréassian, V., 2016. Impact of temporal resolution of inputs on

hydrological model performance: An analysis based on 2400 flood events. J. Hydrol., 538: 454-470. DOI:https://doi.org/10.1016/j.jhydrol.2016.04.016

Fox, J., 2016. Applied Regression Analysis and Generalized Linear Models (Third Edition). SAGE Publications Inc., Califonia.

Gerla, P. J. (2007). Estimating the Effect of Cropland to Prairie Conversion on Peak Storm Run-Off, Restoration Ecology, 15(4), 720-730. https://doi.org/10.1111/j.1526-100X.2007.00284.x

Hall, M.J., 1984. Urban Hydrology. Elsevier Applied Science Publishers, Lodon.

Harrell F.E, Jr. 2001. Regression Modeling Strategies. Springer-Verlag: New York, NY.

Harrod, T.R., Hogan, D.V., 2008. The soils of North Wyke and Rowden, North Wyke Research, North Wyke, Devon.

Hooghoudt, S.B. 1940. General consideration of the problem of field drainage by parallel drains, ditches, watercourses, and channels. Publ. No.7 in the series Contribution to the knowledge of some physical parameters of the soil (titles translated from Dutch). Bodemkundig Instituut, Groningen, The Netherlands.

Huang, J., Kang, Q., Yang, J.X., Jin, P.W., 2017. Multifactor analysis and simulation of the surface runoff and soil infiltration at different slope gradients. IOP Conf. Ser. Earth Environ. Sci., 82: 012019. DOI:https://doi.org/10.1088/1755-1315/82/1/012019

Huang, Y., Bárdossy, A., Zhang, K., 2019. Sensitivity of hydrological models to temporal and spatial resolutions of rainfall data. Hydrol. Earth Syst. Sc., 23(6): 2647-2663. DOI:https://doi.org/10.5194/hess-23-2647-2019

Jansson, P.-E., 1998. Simulation model for soil water and heat conditions. Description of the SOIL model. Division of Agricultural Hydrotechnics Communications 98:2, Swedish University of Agricultural Sciences, Uppsala.

Jeong, J., Kannan, N., Arnold, J., Glick, R., Gosselink, L., Srinivasan, R., 2010. Development and integration of sub-hourly rainfall–runoff modeling capability within a watershed model. Water Resour. Manag., 24(15): 4505-4527. DOI:https://doi.org/10.1007/s11269-010-9670-4

Jones, M. R., H. J. Fowler, C. G. Kilsby and S. Blenkinsop. (2012). An assessment of changes in seasonal and annual extreme rainfall in the UK between 1961 and 2009. International Journal of Climatology, 33(5), 1178-1194, https://doi.org/10.1002/joc.3503.

Kavetski, D., Fenicia, F., Clark, M.P., 2011. Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment. Water Resour. Res., 47: W05501. DOI: https://doi.org/10.1029/2010WR009525

Keesstra, S., Nunes, J., Novara, A., Finger, D., Avelar, D., Kalantari, Z., Cerdà, A., 2018. The superior effect of nature based solutions in land management for enhancing ecosystem services. Sci. Total Environ., 610-611: 997-1009. DOI:https://doi.org/10.1016/j.scitotenv.2017.08.077

Kibet, L.C., Saporito, L.S., Allen, A.L., May, E.B., Kleinman, P.J.A., Hashem, F.M., Bryant, R.B., 2014. A protocol for conducting rainfall simulation to study soil runoff. J. Vis. Exp.(86): e51664. DOI:https://doi.org/10.3791/51664

Ledingham, J., Archer, D., Lewis, E., Fowler, H., Kilsby, C., 2019. Contrasting seasonality of storm rainfall and flood runoff in the UK and some implications for rainfall-runoff methods of flood estimation. Hydrol. Res., 50(5): 1309-1323. DOI:https://doi.org/10.2166/nh.2019.040

Liu, Y., Li, Y., Harris, P., Cardenas, L., Dunn, R.M., Sint, H., Murray, P., Lee, M., Wu, L., 2018. Modelling field scale spatial variation in water run-off, soil moisture, N_2O emissions and herbage biomass of a grazed pasture using the SPACSYS model. Geoderma, 315: 49-58. DOI:https://doi.org/10.1016/j.geoderma.2017.11.029

Liu, Y., Wu, L., Watson, C.A., Baddeley, J.A., Pan, X., Zhang, L., 2013. Modeling biological dinitrogen fixation of field pea with a process-based simulation model. Agron. J., 105(3): 670-678. DOI:https://doi.org/10.2134/agronj2012.0412

Macleod, C.J.A., Humphreys, M.W., Whalley, W.R., Turner, L., Binley, A., Watts, C.W., Skot, L., Joynes, A., Hawkins, S., King, I.P., O'Donovan, S., Haygarth, P.M., 2013. A novel grass hybrid to reduce flood generation in temperate regions. Sci. Rep., 3: 1683. DOI:https://doi.org/10.1038/srep01683

Merz, B., Plate, E.J., 1997. An analysis of the effects of spatial variability of soil and soil moisture on runoff. Water Resour. Res., 33(12): 2909-2922. DOI:https://doi.org/10.1029/97WR02204

Merz, R., Parajka, J., Blöschl, G., 2009. Scale effects in conceptual hydrological modeling. Water Resour. Res., 45: W09405. DOI: https://doi.org/10.1029/2009WR007872

Morison, J.I.L., Matthews, R.B., 2016. Agriculture and forestry climate change impacts summary report. DOI:https://nerc.ukri.org/research/partnerships/ride/lwec/report-cards/agriculture/

Newell-Price, P., Chambers, B., Whittingham, M., 2011. The alleviation of grassland compaction by mechanical soil loosening. Defra Project BD5001 Final Report.

Newell Price, B., Chambers, B., Whittingham, M., 2012. Characterisation of Soil Structural Degradation Under Grassland and Development of Measures to Ameliorate its Impact on Biodiversity and other Soil Functions. Defra Project BD5001 Final Report.

Orr, R., Murray, P., Eyles, C., Blackwell, M., Cardenas, L., Collins, A.L., Dungait, J., Goulding, K., Griffith, B., Gurr, S., Harris, P., Hawkins, J., Misselbrook, T., Rawlings, C., Shepherd, A., Sint, H., Takahashi, T., Tozer, K., Whitmore, A.P., Wu, L., Lee, M., 2016. The North Wyke Farm Platform: effect of temperate grassland farming systems on soil moisture contents, runoff and associated water quality dynamics. Eur. J. Soil Sci., 67(4): 374-385. DOI:https://doi.org/10.1111/ejss.12350

Palmer, R.C., Smith, R.P., 2013. Soil structural degradation in SW England and its impact on surface-water runoff generation. Soil Use Manag., 29(4): 567-575. DOI:https://doi.org/10.1111/sum.12068

Pathak, P., Sudi, R., Wani, S.P., Sahrawat, K.L., 2013. Hydrological behavior of Alfisols and Vertisols in the semi-arid zone: Implications for soil and water management. Agric. Water Manag., 118: 12-21. DOI:https://doi.org/10.1016/j.agwat.2012.11.012

Perego, A., Wu, L., Gerosa, G., Finco, A., Chiazzese, M., Amaducci, S., 2016. Field evaluation combined with modelling analysis to study fertilizer and tillage as factors affecting N_2O emissions: a case study in Po valley (Northern Italy) Agric. Ecosyst. Environ., 225: 72-85. DOI:https://doi.org/10.1016/j.agee.2016.04.003

Pulley, S., Collins, A.L., 2019. Field-based determination of controls on runoff and fine sediment generation from lowland grazing livestock fields. J. Environ. Manage., 249: 109365. DOI:https://doi.org/10.1016/j.jenvman.2019.109365

Regenass, D., L. Schlemmer, E. Jahr and C. Schär. It rains and then? Numerical challenges with the 1D Richards equation in kilometer-resolution land surface modelling. Hydrology and Earth System Sciences. https://doi.org/10.5194/hess-2021-426.

Rust, W., Corstanje, R., Holman, I.P., Milne, A.E., 2014. Detecting land use and land management influences on catchment hydrology by modelling and wavelets. J. Hydrol., 517: 378-389. DOI:https://doi.org/10.1016/j.jhydrol.2014.05.052

Sharma, A., Wasko, C., Lettenmaier, D.P., 2018. If precipitation extremes are increasing, why aren't floods? Water Resour. Res., 54(11): 8545-8551. DOI:https://doi.org/10.1029/2018wr023749

Stevens, A.J., Clarke, D., Nicholls, R.J., 2016. Trends in reported flooding in the UK: 1884–2013. Hydrolog. Sci. J., 61(1): 50-63. DOI:https://doi.org/10.1080/02626667.2014.950581

Verstraten, L., Wasko, C., Ashford, G., Sharma, A., 2019. Sensitivity of Australian roof drainage structures to design rainfall variability and climatic change. Build. Environ., 161: 106230. DOI:https://doi.org/10.1016/j.buildenv.2019.106230

Wasko, C., Sharma, A., Lettenmaier, D.P., 2019. Increases in temperature do not translate to increased flooding. Nat. Commun., 10: 5676. DOI:https://doi.org/10.1038/s41467-019-13612-5

Watts, G., Anderson, M., 2016. Water climate change impacts report card 2016. DOI:https://nerc.ukri.org/research/partnerships/ride/lwec/report-cards/water/

Wu, L., Bingham, I.J., Baddeley, J.A., Watson, C.A., 2009. Modeling plant nitrogen uptake using three-dimensional and one-dimensional root architecture. In: Ma, L., Ahuja, L.R., Bruulsema, T.W. (Eds.), Quantifying and understanding plant nitrogen uptake systems modeling. CRC Press, Boca Raton, FL, pp. 197-218.

Wu, L., Blackwell, M., Dunham, S., Hernández-Allica, J., McGrath, S.P., 2019. Simulation of phosphorus chemistry, uptake and utilisation by winter wheat. Plants, 8(10): 404. DOI:https://doi.org/10.3390/plants8100404

Wu, L., McGechan, M.B., McRoberts, N., Baddeley, J.A., Watson, C.A., 2007. SPACSYS: integration of a 3D root architecture component to carbon, nitrogen and water cycling - model description. Ecol. Model., 200(3-4): 343-359. DOI:https://doi.org/10.1016/j.ecolmodel.2006.08.010

Wu, L., Rees, R.M., Tarsitano, D., Zhang, X., Jones, S.K., Whitmore, A.P., 2015. Simulation of nitrous oxide emissions at field scale using the SPACSYS model. Sci. Total Environ., 530–531(0): 76-86. DOI:http://dx.doi.org/10.1016/j.scitotenv.2015.05.064

Wu, L., Zhang, X., Griffith, B.A., Misselbrook, T., 2016. Sustainable grassland systems: A modelling perspective based on the North Wyke Farm Platform. Eur. J. Soil Sci., 67(4): 397-408. DOI:https://doi.org/10.1111/ejss.12304

Zehe, E., Graeff, T., Morgner, M., Bauer, A., Bronstert, A., 2010. Plot and field scale soil moisture dynamics and subsurface wetness control on runoff generation in a headwater in the Ore Mountains. Hydrol. Earth Syst. Sci., 14(6): 873-889. DOI:https://doi.org/10.5194/hess-14-873-2010

Ziadat, F. M. and A. Y. Taimeh. (2013). Effect of Rainfall Intensity, Slope, Land Use and Antecedent Soil Moisture on Soil Erosion in an Arid Environment, Land Degradation & Development, 24(6), 582-590. https://doi.org/10.1002/ldr.2239.

Zhang, X., Sun, N., Wu, L., Xu, M., Bingham, I.J., Li, Z., 2016. Effects of enhancing soil organic carbon sequestration in the topsoil by fertilization on crop productivity and stability: evidence from long-term experiments with wheat-maize cropping systems in China. Sci. Total Environ., 562: 247-259. DOI:https://doi.org/10.1016/j.scitotenv.2016.03.193

4. Adjusting for conditional bias in process model simulations of hydrological extremes: an experiment using the North Wyke Farm Platform

Stelian Curceac[1*], Peter M. Atkinson[2,3,4], Alice Milne[5], Lianhai Wu[1], Paul Harris[1]

[1] Rothamsted Research, Department of Sustainable Agriculture Sciences, North Wyke EX20 2SB, Devon, UK.

[2]Lancaster Environment Centre, Lancaster University, Bailrigg, Lancaster LA1 4YQ, UK.

[3]Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

[4]State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China.

[5]Rothamsted Research, Department of Sustainable Agriculture Sciences, Harpenden AL5 2JQ, UK

***Correspondence:**

Stelian Curceac

stelian.curceac@rothamsted.ac.uk

Published in Frontiers in Artificial Intelligence, AI in Food, Agriculture and Water, Research Topic: Machine Learning for Water Resources

4.1 Abstract

Peak flow events can lead to flooding which can have negative impacts on human life and ecosystem services. Therefore, accurate forecasting of such peak flows is important. Physically-based process models are commonly used to simulate water flow, but they often under-predict peak events (i.e., are conditionally biased), undermining their suitability for use in flood forecasting. In this research, we explored methods to increase the accuracy of peak flow simulations from a process-based model by combining the model's output with: (a) a semi-parametric conditional extreme model and (b) an extreme learning machine model. The proposed 3-model hybrid approach was evaluated using fine temporal resolution water flow data from a sub-catchment of the North Wyke Farm Platform, a grassland research station in south-west England, UK. The hybrid model was assessed objectively against its simpler constituent models using a jackknife evaluation procedure with several error and agreement indices. The proposed hybrid approach was better able to capture the dynamics of the flow process and, thereby, increase prediction accuracy of the peak flow events.

Keywords: peak flow, conditional extreme model, extreme learning machine, process-based model, hybrid, grassland agriculture.

4.2 Introduction

In the UK, the estimated yearly cost of damages caused by floods is over £1 billion (Collet et al., 2017). Accurate and reliable forecasting of extreme flow events is crucial for planning and implementing measures to mitigate their effects and so protect lives, properties and services. The magnitude and frequency of floods is likely to increase as a result of climate change (Bates

et al., 2008; Field et al., 2012; Kundzewicz et al., 2007) and this could push ecosystems beyond the threshold of normal disturbance (Thibault & Brown, 2008). Increased runoff and flooding intensify erosion and result in higher sediment and nutrient losses that can lead to soil degradation and high concentrations of pollutants in water courses (Bouraoui et al., 2004).

Over recent decades, different approaches have been proposed for more accurate modelling and forecasting of peak flows with reduced uncertainty. The two main methods of modelling hydrological variables are physically-based models and statistical models. However, there is an increasing trend towards combining these approaches in hybrid models. One of the most common ways to do this is to post-process statistically an ensemble of forecasts from process-based models (e.g., Cloke and Pappenberger, 2009; Li et al., 2017). Bayesian methods using climate indices (Bradley et al., 2015), stochastic data-driven methods on wavelet decomposed series (Quilty et al., 2019), Bayesian model averaging (Raftery et al., 2005), extended logistic regression (Roulin and Vannitsem, 2011), quantile regression (López López et al., 2014), bias correction (Li et al., 2019) and nearest neighbor resampling for uncertainty estimation (Sikorska et al., 2015) are among the many post-processing techniques described in the literature. Examples of combining a process-based model with more than one statistical or machine learning model can be found in Bogner et al. (2017), Papacharalampous et al. (2019) and Tyralis et al. (2019). The usefulness of combining deterministic and stochastic models (Box and Jenkins, 1976) in real-time flood forecasting was reported by Toth et al. (1999), while the performance of various post-processing techniques according to the level of flow was investigated in Bogner et al. (2016) and Papacharalampous et al. (2019). Hybrid methods for water flow (streamflow) forecasting also include the combination of classical statistical methods with more data-driven, machine-learning methods such as artificial neural networks

(ANNs) (Chen et al., 2018; Yaseen et al., 2016; Zhou et al., 2018), discrete wavelet transforms and support vector machines (Kisi and Cimen, 2011), and coupling ANNs with autoregressive techniques (Fathian et al., 2019). The effect of catchment characteristics on the predictive performance of two different statistical models was discussed in Dogulu et al. (2015).

Hydrological process-based models (PBMs) are traditionally used for streamflow modelling and forecasting, where under-prediction of peak flows is a common issue (e.g., Lane et al., 2019; Wijayarathne and Coulibaly, 2020). The PBM performance can suffer from uncertainty due to both random and systematic errors. Both random and systematic errors can arise in the estimated model parameters and measured input variables. However, of particular interest is a type of systematic error (or bias) called conditional bias that depends on flow magnitude. That is, the structure and parameters of the model can generalise the outputs leading to conditional bias, specifically under-prediction of large values and over-prediction of small values; an effect similar in nature to that of having a support that is larger than ideal. Alternatively, data-driven methods may be used, especially when the initial conditions and the parameters of the physical model are difficult to estimate or when the length and/or quality of the data are insufficient for a reliable model calibration.

In this research, we explored combining statistical and machine learning techniques with flow simulations obtained from a PBM to increase the accuracy of forecasting peak flow events. Specifically, we considered the semi-parametric, conditional extreme model (CEM) of Heffernan and Tawn (2004) (a statistical model) and the extreme learning machine (ELM) of Huang et al. (2006) (a machine learning model). The proposed approach is considered a generic solution for enhancing any given hydrological PBM.

The CEM is appropriate for describing the probability that one or multiple variables are extreme and has been applied widely for flood risk analysis (Mendes and Pericchi, 2009; Lamb et al., 2010; Keef et al., 2013; Zheng et al., 2014). A significant property of the CEM is that it is flexible in modelling different dependence structures, such as the dependence of different variables at the same site or the dependence of the same variable at different sites. A key assumption of the application of the CEM is that the extremes of each variable must be independent and, consequently, cannot be used to model peak flow events that have a duration of several consecutive days and, therefore, exhibit temporal dependence. For this reason, the maximum flow during each event was modelled using the CEM while all other peaks were modelled using the ELM (and, thus, a 3-model rather than a 2-model hybrid is proposed).

The ELM model is ANN-based and has been used in various areas of water resources engineering, with a recent focus on water flow (see Yaseen et al., 2019 for an extensive review). In this context, it has been shown to increase accuracy and reduce computational time compared to commonly used benchmark models (Lima et al., 2015) and to other ANN models (Deo and Şahin, 2016).

The resultant 3-model hybrid was evaluated empirically using measured flow data from a sub-catchment of the North Wyke Farm Platform, a grassland research facility in south-west England (Orr et al., 2016). To our knowledge, no study to-date has used the CEM and the ELM to improve the simulation of peak flow events obtained from a PBM, or in which they are combined. The proposed methodology builds on the modelled dependence structure between measured and PBM-simulated peak flow events and uses this relationship to obtain a more accurate representation of these events.

4.3 Methods

This section presents a general description of the CEM (Heffernan and Tawn, 2004) and the ELM (Huang et al., 2006) and explains how they can be applied to peak flow events obtained from a chosen PBM (described in Section 4.4.2) in a hybrid context. The flow threshold, above which the simulated and the observed data are considered as possible peaks, is determined based on Generalised Pareto Distribution (GPD) stability plots of the PBM simulated values (Curceac et al., 2020). The performance of the proposed hybrid approach is evaluated using a jackknife procedure and by calculating several error and agreement indices.

4.3.1 Generalised Pareto Distribution (GPD)

We characterise peak flow events by fitting the GP distribution to the extreme flow above a certain threshold. The cumulative distribution function (CDF) of the independent and identically distributed (iid) excesses over an appropriately high threshold u for the GPD is:

$$G(x) = \Pr(X - u < x | X > u) = \begin{cases} 1 - \left(1 + \dfrac{\xi(x-u)}{\sigma}\right)^{-\frac{1}{\xi}}, \xi \neq 0 \\ 1 - e^{\left(-\frac{x-u}{\sigma}\right)}, \quad \xi = 0 \end{cases} \qquad 4\text{-}1$$

where x, for this study, is the peak flow in mm d^{-1}, u is the location parameter, σ is the scale parameter and ξ is the shape parameter. The value of the shape parameter defines the type of distribution from the GPD family; that is, $\xi = 0$ refers to the exponential distribution, the distribution has an upper bound of $u - \sigma/\xi$ when $\xi < 0$ and has no upper limit when $\xi \geq 0$.

The first step in modelling the exceedances is to select a threshold over which peaks in flow are considered extreme. The next step is to ensure that the peaks above it are independent (so as to conform with iid) and estimate the scale and shape parameters. The selection of the

threshold is a crucial step in GPD extreme value analysis and is basically a trade-off between bias (low threshold-large sample size) and variance (high threshold-small sample size).

The flow threshold in this research was selected based on the simulated flow from the study's PBM using an automated threshold stability method (Curceac et al., 2020) (Section 4.3.2) and the same threshold was used for the measured flow data. The GP model was fitted initially independently to the simulated and observed peak flows and the conditional dependence structure between them was estimated using the CEM (Section 4.3.3).

4.3.2 GPD Threshold Selection

If the GPD is an appropriate model for the excesses above a threshold u, then for all larger thresholds $u^* > u$ it will also be suitable with the shape parameter being relatively constant (Coles, 2001; Scarrott & MacDonald, 2012). That is, it is the approximately linear and horizontal segment on a plot of shape parameter against threshold. This does not apply for the scale parameter σ_{u^*}, which changes with the threshold $\sigma_{u^*} = \sigma_u + \xi(u^* - u)$. However, the modified scale parameter $\sigma_1 = \sigma_{u^*} - \xi u$ remains relatively constant. Therefore, following Curceac et al. (2020), we fitted a cubic smoothing spline to this plot and calculated the rate of change at each of m consecutive steps. The cubic smoothing spline estimate $\hat{f}$ of a function f in the model $Y_i = f(x_i) + \varepsilon_i$, is defined as the minimizer of $\sum_{i=1}^{n}\{Y_i - \hat{f}(x_i)\}^2 + \lambda \int \hat{f}''(x)^2 dx$, where λ is the smoothing parameter. The minimum change rate locates the part of the plot where the shape and the modified scale parameters reach a plateau.

4.3.3 Conditional Extreme Model (CEM)

For a continuous *d*-dimensional vector variable $X = (X_1, \dots, X_d)$ with unknown distribution function $F(x)$, the CEM describes the distribution function of X when it is extreme in at least

one component. In other words, it describes the conditional distribution of $X_{-i}|X_i > u_{X_i}$, where X_{-i} is the vector variable X without the component X_i.

After estimating the marginal distribution of each $X_i, i = 1, \dots, d$ (Section 4.3.1), and before estimating the extremal dependence, the variables are transformed so that they follow the same distribution. This process is called marginal standardization and is used to distinguish the marginal behaviour from the dependence structure (Drees and Janßen, 2017). The data can be transformed to either Gumbel margins to describe the positive dependence or to a Laplace marginal distribution which, due to its exponential tail and symmetry, captures both positive and negative dependence (Keef et al., 2013). The initial vector variable X is, therefore, transformed as:

$$f(x) = \begin{cases} \log\{2F_{X_i}(X_i)\}, & X_i < F_{X_i}^{-1}(0.5) \\ -\log\{2[1 - 2F_{X_i}(X_i)]\}, & X_i \geq F_{X_i}^{-1}(0.5) \end{cases} \quad \text{4-2}$$

where $F_{X_i}^{-1}$ is the inverse cumulative distribution function of X_i. The resulting vector variable $Y = (Y_1, \dots, Y_d)$, therefore, has Laplace margins with:

$$\Pr(Y_i \leq y) = F_{Y_i}(y) = \begin{cases} \frac{1}{2}\exp(y), & y < 0 \\ 1 - \frac{1}{2}\exp(-y), & y \geq 0 \end{cases} \quad \text{4-3}$$

The dependence model considers the asymptotics of the conditional distribution $\Pr(Y_{-i} \leq y_{-i}|Y_i = y_i)$, where for $y_i \to \infty$, the increase of y_{-i} must result in non-degenerate margins. For this, assume the normalizing functions $a_{|i}(y_i)$ and $b_{|i}(y_i)$, that have the same dimension as Y_{-i} and for which:

$$\lim_{y_i \to \infty} \left[\Pr\left\{ \frac{Y_{-i} - \mathrm{a}_{|i}(y_i)}{b_{|i}(y_i)} \leq z_{|i} \middle| Y_i = y_i \right\} \right] = G_{|i}(z_{|i}) \qquad 4\text{-}4$$

where the limit distribution $G_{|i}$ has non-degenerate marginals $G_{j|i}$ for all $j \neq i$. Therefore, the random variable $Z_{|i} = \frac{Y_{-i} - \mathrm{a}_{|i}(y_i)}{b_{|i}(y_i)}$ is independent of $Y_i > u_{Y_i}$ and has distribution function $G_{|i}$. The location $a_{|i}(y_i)$ and scale $b_{|i}(y_i)$ functions are given by $a_{|i}(y_i) = \alpha_{|i} y_i$ and $b_{|i}(y_i) = {y_i}^{\beta_{|i}}$ where the vector constants $\alpha_{|i}$ and $\beta_{|i}$ take values of $\alpha_{j|i} \in [-1,1]$ and $\beta_{j|i} \in (-\infty, 1)$, respectively, for all $j \neq i$. Finally, the dependence structure is described by the multivariate semi-parametric regression model:

$$Y_{-i} = \alpha_{|i} y_i + {y_i}^{\beta_{|i}} Z_{|i} \text{ for } Y_i = y_i > u_{Y_i},\ i = 1, \ldots, d. \qquad 4\text{-}5$$

The above equation expresses the behaviour of the vector variable Y, excluding the element of Y_i when it takes a large value. The dependence between the variables Y_i and Y_j is explained by the constant $\alpha_{j|i}$. Positive values indicate a positive relationship. The constant $\beta_{j|i}$ incorporates the changes in the variability of Y_j as Y_i increases. Details on estimating the dependence parameters are given in Heffernan and Tawn (2004) and Keef et al. (2013).

To obtain randomly generated samples of $X|X_i > u_{X_i}$, we adopted the following procedure. Initially, samples of Y_i from the Laplace distribution are simulated conditional on it exceeding its cumulative probability corresponding to $F_{X_i}(u_{X_i})$. Similarly, samples of random observations of $Z_{|i}$ are drawn from its estimated distribution $\hat{G}_{|i}$. Then, using the semi-parametric model, we obtain $Y_{-i} = \hat{\alpha}_{|i} y_i + {y_i}^{\hat{\beta}_{|i}} Z_{|i}$ and transform the vector $Y = (Y_{-i}, Y_i)$ to the originally distributed $X = (X_{-i}, X_i)$ by the inverse transformation.

4.3.4 Extreme Learning Machine (ELM)

The ELM is a data-driven method developed by Huang et al. (2006) that has been used effectively for streamflow forecasting (e.g., Deo and Şahin, 2016; Yaseen et al., 2016). Compared to other common ANN techniques, it has the advantages of fast learning speed and is characterised by improved performance in terms of commonly encountered problems, such as over-fitting and the effect of local minima. The model has a three-layer structure with one input, one hidden and a single output layer and can be expressed mathematically as:

$$\sum_{i=1}^{\Lambda} B_i h_i(m_i \cdot x_t + n_i) = z_t \qquad 4\text{-}6$$

where Λ is the total number of nodes, B are the estimated weights between the nodes of the hidden and output layers, and $h(m, n, x)$ is the activation function with weights $m_i \in \Re^d$, biases $n_i \in \Re$ and the explanatory variable of the training dataset $x_t \in \Re^d$. Here, i and d denote the index of a specific hidden neuron (HN) and the number of input neurons, respectively, and Z is the model output.

Initially, the ELM model selects the input weights and hidden layer biases at random, and then calculates the output weights using a least squares method instead of adjusting them iteratively (see Chen et al. 2018 for details). Once the output weights $\hat{B}$ have been estimated, forecasts are obtained by substituting the training dataset x_t with the testing one. The number of HNs in the hidden layer and the activation function are the only parameters that need to be pre-defined. The optimal number of HNs is a trade-off between generalization ability and network complexity. A highly complex model with too many HNs can lead to over-fitting, whereas a decreased number of HNs can result in a model that is too simple to capture

non-linear relationships. The optimal number of HNs is problem-dependent and is frequently determined empirically (Huang et al., 2006; Sun et al., 2008). In this research, the number of HNs was increased iteratively from 1 to 100 and the network structure that provided the smallest RMSE of the training procedure was selected.

4.3.5 Application and Evaluation

A jackknife evaluation procedure (Miller, 1964; Shao and Tu, 1995) was applied to assess the performance of the proposed hybrid approach. It is a leave-one-out resampling technique without random replacement where one observation or a fixed subset of the dataset is omitted iteratively. For a sample containing n data points, an analysis is performed n times each on $n-1$ data points. The main strengths of the jackknife method are that model accuracy is independent of the calibration data and the loss in the sample data information is minimal (McCuen, 2005).

As stated previously, peak events are defined as flow above a certain threshold of the PBM simulated data. At each iteration, one peak flow event (measured and simulated) was left out of the dataset. This event constitutes the testing dataset and the rest of the data the training dataset, and the CEM and the ELM were fitted to the latter. The dependence behavior of measured peaks conditional on the PBM simulated, above a certain threshold, was configured by the CEM. From the fitted CEM, 50,000 stochastic simulations were obtained for both the observed X_j (pseudo-observations) and the PBM simulated X_i variables (pseudo-PBM simulated). From the total set of random simulations of the conditioning variable X_i, the ones with the smallest difference (≤ 0.1) from the maximum PBM simulated peak of the testing sample, which was left out of the training dataset, were considered. As CEM provides pairs of simulated data according to their dependence structure, the corresponding random

simulations of X_j (pseudo-observations) were then obtained. By calculating their median value, a forecast of the maximum flow during an event was obtained and compared to the maximum measured and PBM simulated peak excess of the testing dataset.

The ELM model was trained using PBM simulated data as inputs and measured data as outputs of the training dataset. Based on the trained ELM model, flow forecasts were then obtained using the PBM simulated flow of the testing sample as explanatory variable, except for the maximum. Consequently, peaks smaller than the cluster maxima were forecasted by the ELM and the CEM was used only to forecast maximum flows. The application of the ELM model alone on all the peaks was also performed in experimentation and its performance compared to the CEM for the maximum flows. At the next iteration, a different peak flow event was omitted from the training dataset for testing purposes and the same process was repeated for all peaks.

This procedure was performed initially for peaks above the threshold that corresponds to the start of the region of stability of shape and modified scale parameters. However, in order to investigate the effect of threshold selection on the proposed methodology, the above-mentioned procedure was repeated for different thresholds. The considered thresholds were set as a range from the minimum that resulted from the application of threshold stability method, up to the 95th quantile of the PBM simulated flow. Higher thresholds resulted in data scarcity that did not allow the models to be fitted satisfactorily. All the above-mentioned steps are presented diagrammatically in Figure 4-1.

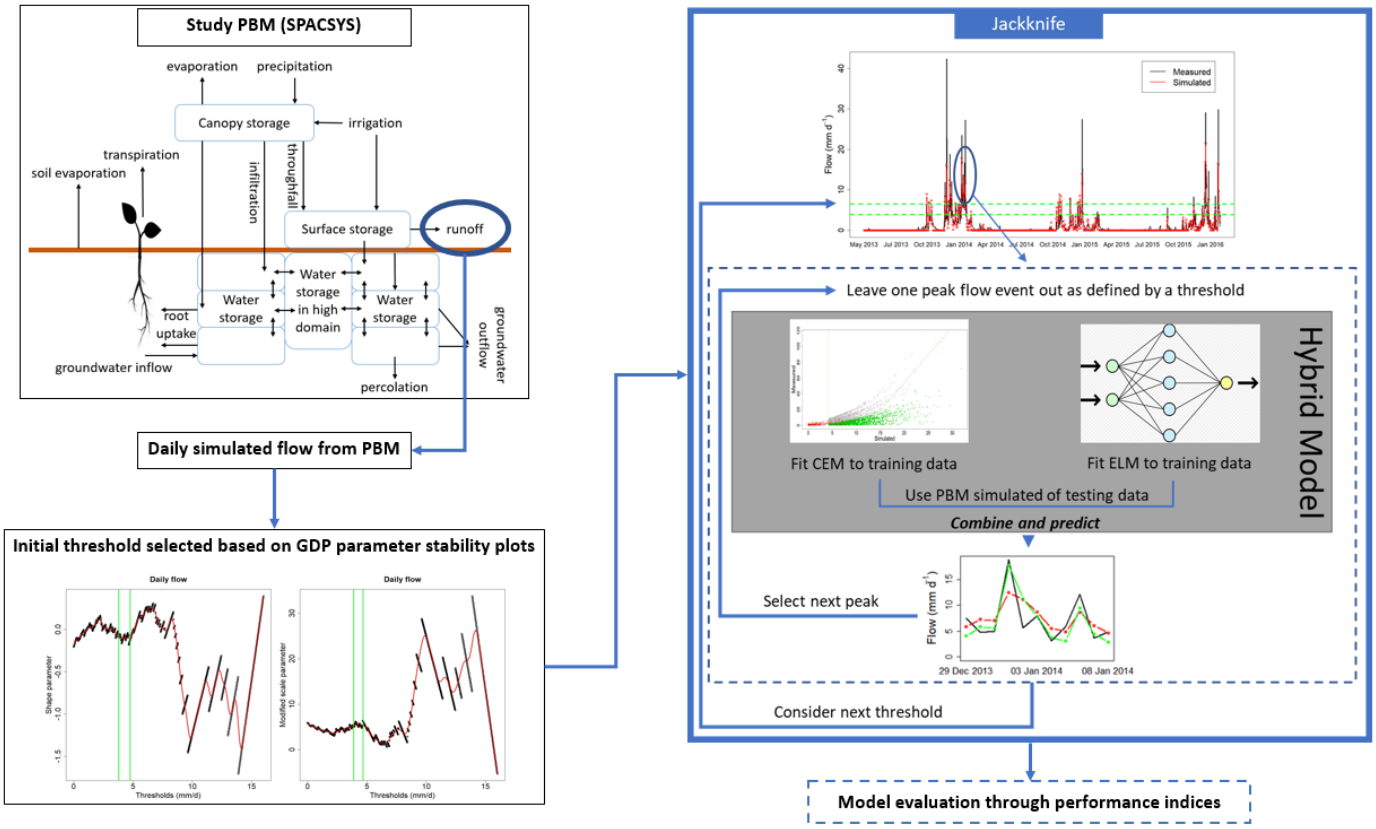


Figure 4-1: Schematic of the proposed methodology.

To assess the accuracy of the peak flow forecasts for each threshold, a set of indices was calculated. More specifically, the mean absolute error (MAE), the normalized root mean square error (NRMSE), the percentage BIAS (PBIAS), the Nash-Sutcliffe efficiency (NSE), the index of agreement (*d*) and the Kling-Gupta Efficiency (KGE) were computed using the following equations:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{z}_i - z_i| \qquad 4\text{-}7$$

$$\mathrm{NRMSE} = 100\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{z}_i - z_i)^2}}{z_{\max} - z_{\min}} \qquad 4\text{-}8$$

$$\mathrm{PBIAS} = 100\frac{\sum_{i=1}^{N}(\hat{z}_i - z_i)}{\sum_{i=1}^{N} z_i} \quad 4\text{-}9$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(\hat{z}_i - z_i)^2}{\sum_{i=1}^{N}(z_i - \bar{z}_i)^2} \quad 4\text{-}10$$

$$d = 1 - \frac{\sum_{i=1}^{N}(\hat{z}_i - z_i)^2}{\sum_{i=1}^{N}(|\hat{z}_i - \bar{z}_i| + |z_i - \bar{z}_i|)^2} \quad 4\text{-}11$$

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{\hat{z}}}{\sigma_z} - 1\right)^2 + \left(\frac{\bar{\hat{z}}}{\bar{z}} - 1\right)^2} \quad 4\text{-}12$$

where $\hat{z}_i$ are the simulated (or predicted) values, z_i are the measurements (or observed values), $\bar{z}_i$ is the mean of the measured values, r is the Pearson product-moment correlation coefficient (between $\hat{z}_i$ and z_i) and σ is the standard deviation. The optimal value of the error indices (MAE, NRMSE and PBIAS) is zero and the smaller are the values, the more accurate are the simulations. NSE (Nash and Sutcliffe, 1970) takes values from $-\infty$ to 1, where one corresponds to a perfect match between simulated and measured values, zero indicates that model simulations are as accurate as the mean of the measured values and a negative value indicates that the mean of the measured values is a more accurate predictor than the model. The index of agreement, d is defined in the range of zero to one, where again one represents the perfect model and zero no agreement at all. KGE incorporates r, the ratio between the means of the measurements and the simulations, and the variability ratio. KGE takes the same value range as NSE.

4.4 Study Site and Data

4.4.1 Study site

The flow discharge data used in this research were measured at the North Wyke Farm Platform (NWFP). The NWFP is a farm-scale experiment established in 2010 in the southwest of England (50°46'10"N, 3°54'05"W) to support research into sustainable grassland livestock systems (Orr et al., 2016). The platform comprises three independent small farms, each 21 ha in size. Each farm is divided into five sub-catchments, with some sub-catchments consisting of more than one field. The platform monitors routinely water run-off and water chemistry in each of the 15 sub-catchments, together with other primary data collections (e.g. greenhouse gas emissions) so that each farming system can be evaluated according to its level of sustainability (Takahashi et al., 2018). For the period 1985-2015, the average annual temperature at North Wyke ranges from 6.8 to 13.4 °C and the average annual rainfall is 1033 mm. The platform has an altitude range of 120–180 m above sea level. Soil texture consists of a slightly stony clay loam topsoil (about 36% clay) above a mottled stony clay (about 60% clay). The subsoil is impermeable to water and during rain events most of the excess water moves by surface and sub-surface lateral flow towards the drainage system described below.

Each of the 15 sub-catchments (inset in Figure 4-2) are hydrologically isolated through a combination of topography and a network of French drains (800-mm deep trenches) which ensure that the total runoff is channelled to instrumented flumes, measuring water discharge and its chemistry with a 15 minute temporal frequency since October 2012. The runoff from each sub-catchment is measured through a combination of primary and secondary flow devices. The primary devices are H-type flumes (TRACOM Inc., Georgia, USA) with capacity designed for a 1-in-50-year storm event (in respect of data preceding 2010). The specific

design of the H-type flume facilitates the accurate measurement of both low and high flows and is relatively self-cleaning since it allows the ready passage of sediment and particulate matter. A secondary flow measurement device (OTT hydromet, Loveland, CO., USA) is used to measure the water height within the flume and convert it to discharge rate using flume-specific formulas which depend on water height. The flow is generated only from rainfall as the fields are not irrigated. Each sub-catchment also monitors precipitation and soil moisture every 15 minutes. More detailed description about the NWFP can be found in Section 1.7.

Platform data acquired from October 2011 to July 2013, represent a baseline period where all farm fields were categorized as permanent pasture and received identical rates of inorganic fertilizers and farmyard manure. From July 2013 to July 2015, two of the three farms entered a transition phase and were ploughed and reseeded progressively with different types of pasture; specifically, a mixture of white clover and high sugar perennial ryegrass, and sugar perennial ryegrass only. Thus, two farms entered fully a post-baseline period in July 2015.

For this research, we used flow discharge (from April 2013 to February 2016) measured at sub-catchment 6 of the permanent pasture farm (Figure 4-2), which consists of a single field (Golden Rove). This field was chosen because, as part of the permanent pasture farm, it would not have been ploughed and reseeded during the period of study (which would affect various processes, such as runoff).

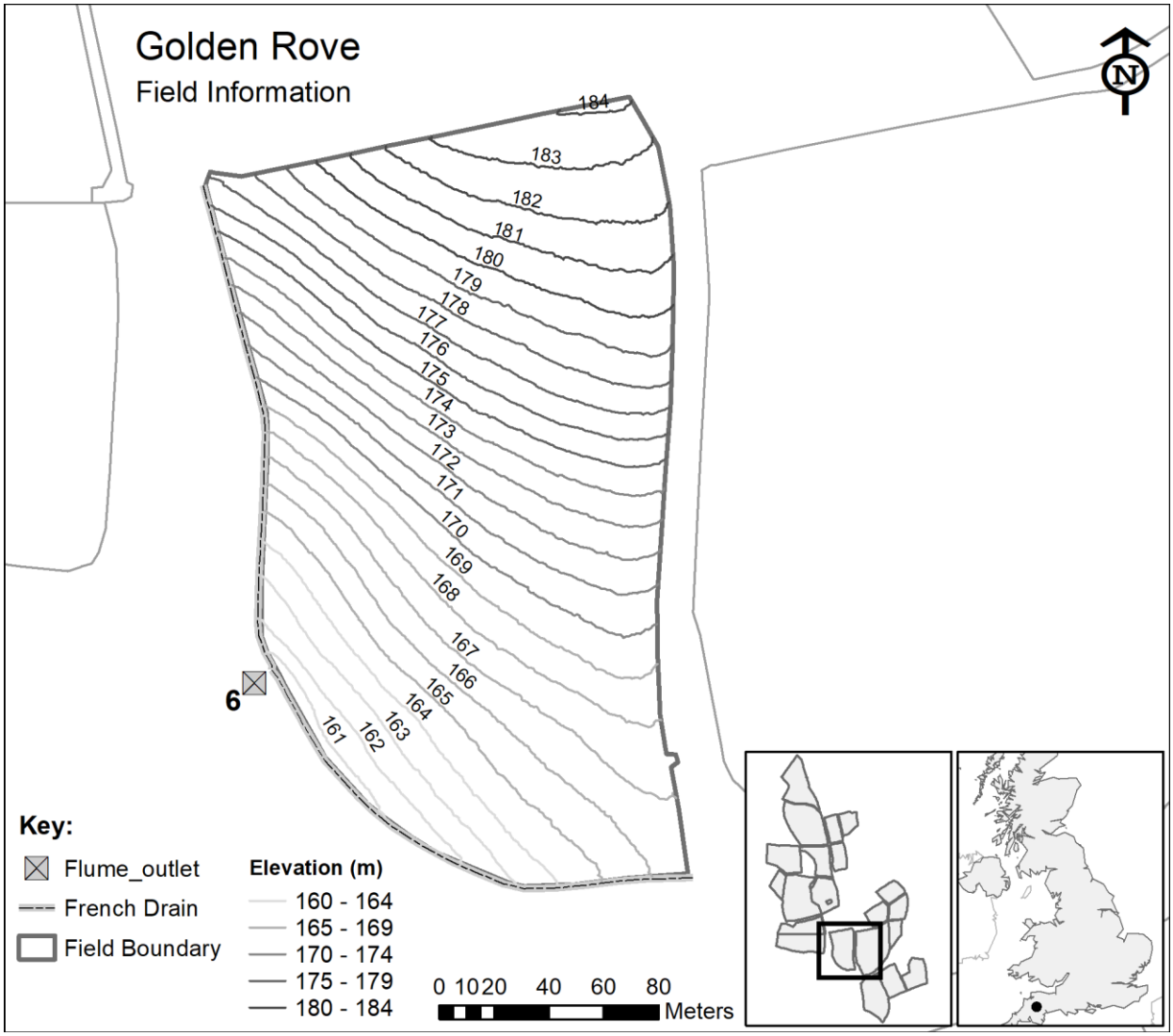


Figure 4-2: Details of the sub-catchment selected for this research from the total of 15 sub-catchments within the North Wyke Farm Platform. Rain gauge location given in Figure 1-2.

4.4.2 Choice of process-based model (PBM)

For this research, we used the 'SPACSYS' model to simulate the flow discharge for sub-catchment 6 of the NWFP over the period of interest. The SPACSYS model is a process-based, field-scale model which simulates key agricultural processes such as plant growth and development, soil carbon (C) and nitrogen (N) cycling, water dynamics and heat transformation (Wu et al., 2007) (see Figure 4-1). The main processes concerning plant growth

are assimilation, respiration, water and N uptake, partitioning of photosynthate and N, N-fixation for legume plants and root growth. The Richards' equation for water potential is used in SPACSYS to simulate water redistribution in a soil profile. Site-specific input data for the simulations include daily weather variables from the North Wyke site, soil properties, field and grass management (e.g., fertiliser application dates and composition, reseeding, grazing and cutting dates), and initialization of the state variables (standing biomass and root distribution, soil water and temperature distribution). Previous simulations of water runoff, soil moisture and other agricultural processes for sub-catchment 6 of the NWFP using SPACSYS can be found in Liu et al. (2018), where a detailed explanation on the SPACSYS calibration is given.

4.5 Results

4.5.1 Comparison of measured flow data with PBM simulations

The plotted time-series of measured and PBM simulated flow (Figure 4-3), shows that the simulation appears to capture well the general behaviour of the process at low flows. However, it tends to under-predict the high flows and over-predict the medium ones. This is confirmed by the corresponding scatterplot (Figure 4-4) and regression line where many values in the range 5-10 mm d^{-1} are below the 1-to-1 line and, thus, the simulated flow is greater than that measured. This results in the regression line also being below the 1-to-1 line although the highest values are above it. The estimated intercept is $a = 0.183$, the estimated slope coefficient $b = 0.883$ and the $R^2 = 0.56$. A non-linear locally weighted regression fit (i.e. a Loess smoother, see Cleveland, 1979), to the measured and simulated data is also given to help illustrate this behaviour.

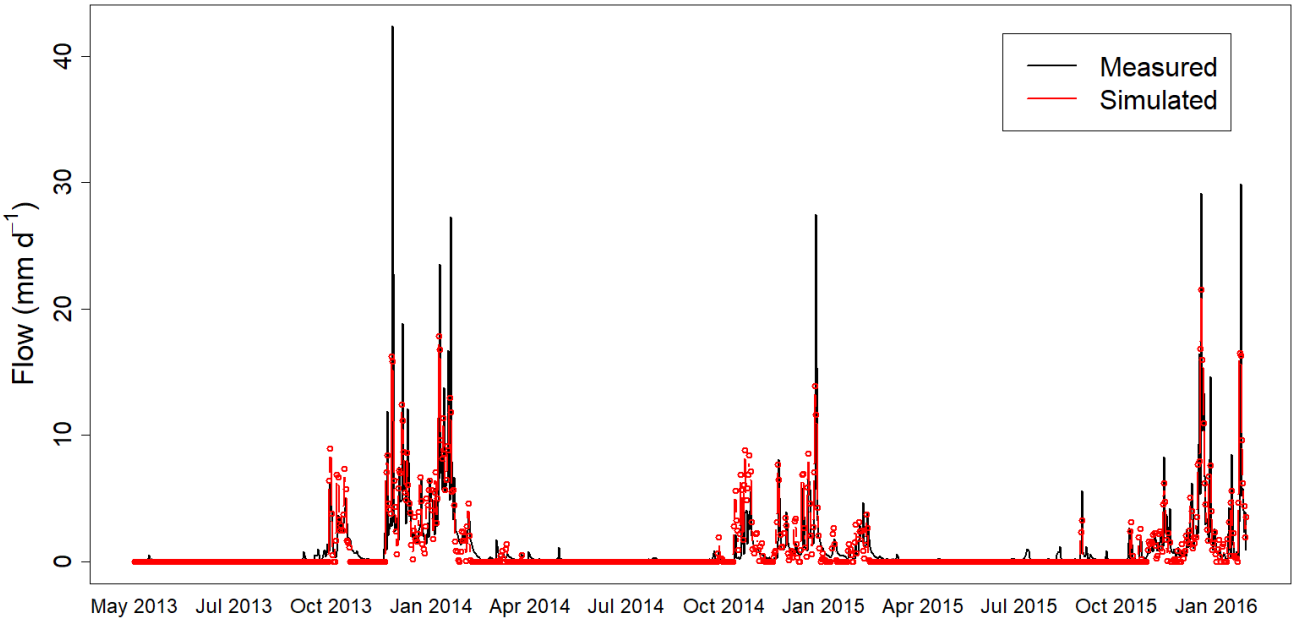


Figure 4-3: Time-series of measurements and PBM simulation of mean daily flow (mm d^{-1}) at the study site from May 2013 to February 2016.

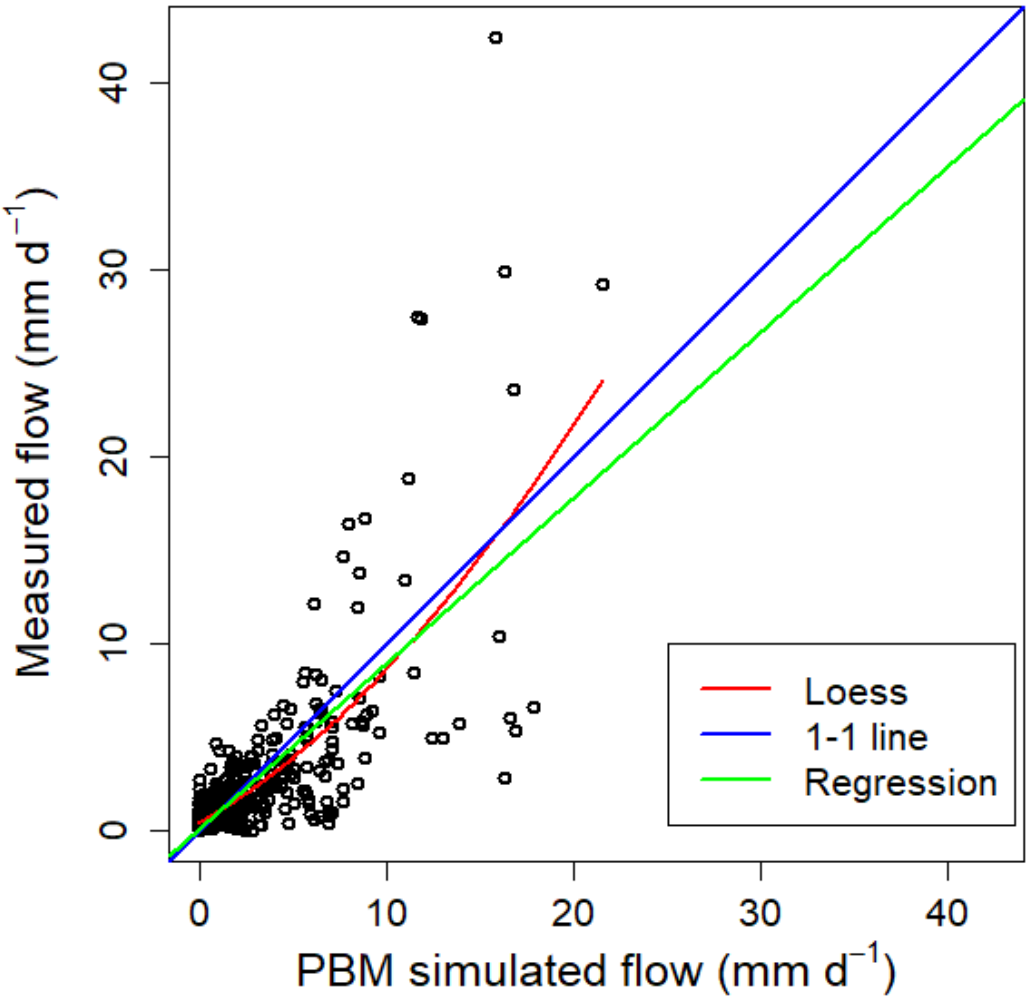


Figure 4-4: Scatterplot of daily measured against daily PBM simulated flow (mm d^{-1}) at the study site. The scatterplot is shown with the ideal 1:1 line, a Loess smoother fit and a regression line.

4.5.2 Threshold selection

The shape and modified scale parameters estimated using the method of Curceac et al. (2020) indicated very similar threshold choices, in regions where the parameters remained relatively stable for increasing threshold candidates (Figure 4-5). The minimum threshold according to the shape parameter is 3.96 mm d^{-1} and according to the modified scale parameter, 3.88 mm d^{-1}. These thresholds were estimated based on the PBM simulated flow (as described above), and the same thresholds were used for the observed peaks. Diagnostics, such as QQ plots of the empirical and modelled distributions (not presented), indicated that the GPD provides a

good fit to the excesses and can model satisfactorily the peaks above the threshold of 3.88 mm d^{-1}, which was eventually selected. The range of thresholds above which the models where applied, was set from 3.88 mm d^{-1} up to 6.41 mm d^{-1}, with the maximum corresponding to the 95th quantile of the PBM simulated flow.

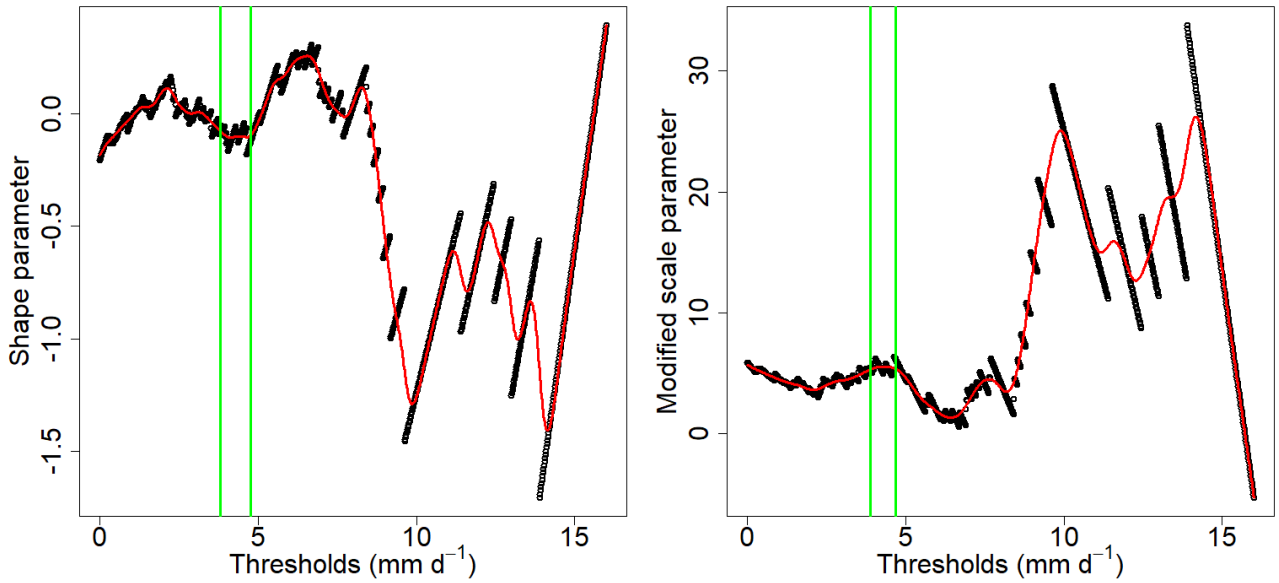


Figure 4-5: Shape and modified scale parameters for different threshold candidates applied to the PBM simulated daily flow. The red lines are the fitted splines and the green vertical lines specify the selected region of stability.

4.5.3 Conditional Extreme Model (CEM) Fit

The diagnostics of the extreme dependence model (CEM) show a satisfactory fit (Figure 4-6). As stated in Section 4.3.3, one of the main assumptions of the model is that the residuals Z are independent of the conditioning variable (in this case, the PBM simulations). The pattern of both the initial and absolute values of the normalized residuals conforms approximately to a uniform distribution with no distinct pattern in the location or scatter of these residuals with the conditioning PBM simulations. The slight trend in the residuals Z for the lowest peaks of the conditioning variable might indicate that a higher threshold should be considered. The

fitted quantiles of the conditional distribution of the dependent variable (measured data) conditional on the PBM simulated data (Figure 4-6, bottom) shows a good agreement between the data and the fitted quantiles, which capture the whole range of the scatter. Histograms of the scale and shape parameters (Figure 4-7) show that the measured and PBM simulated peaks have similar scale characteristics. However, the distribution of the measured peaks has a considerably heavier tail ($\xi_{obs} > \xi_{sims}$). The CEM simulated values of the dependent variable (measured data) along with the values of the conditional variable (PBM simulated data) (Figure 4-8) were obtained using the CEM with estimated dependence parameters of α = 0.44 and β = 0.59. These parameters confirm that there is a positive dependence between the measured and the PBM simulated data, and that the measured data increase in variability as the values of the PBM simulations increase.

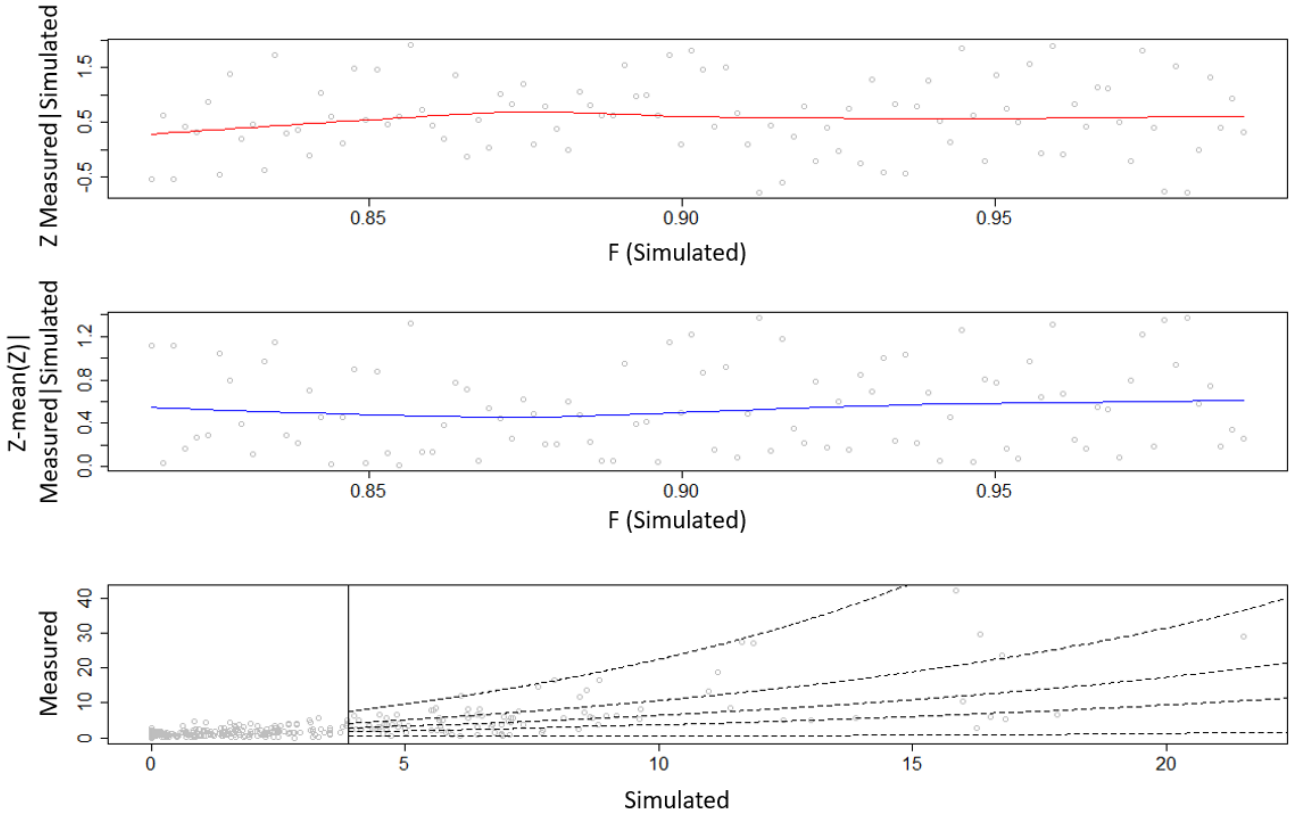


Figure 4-6: Diagnostic plots for the fitted extreme dependence model (CEM): (top) scatterplot of the residuals Z against the conditioning PBM simulated data with a Loess curve (in red) for the local

mean values; (middle) absolute of the normalized residuals Z against the conditioning PBM simulated data with a Loess curve (in blue); (bottom) scatterplot of measured versus PBM simulated data, with the fitted quantiles of the distribution of measured data conditional on PBM simulated data (dashed lines).

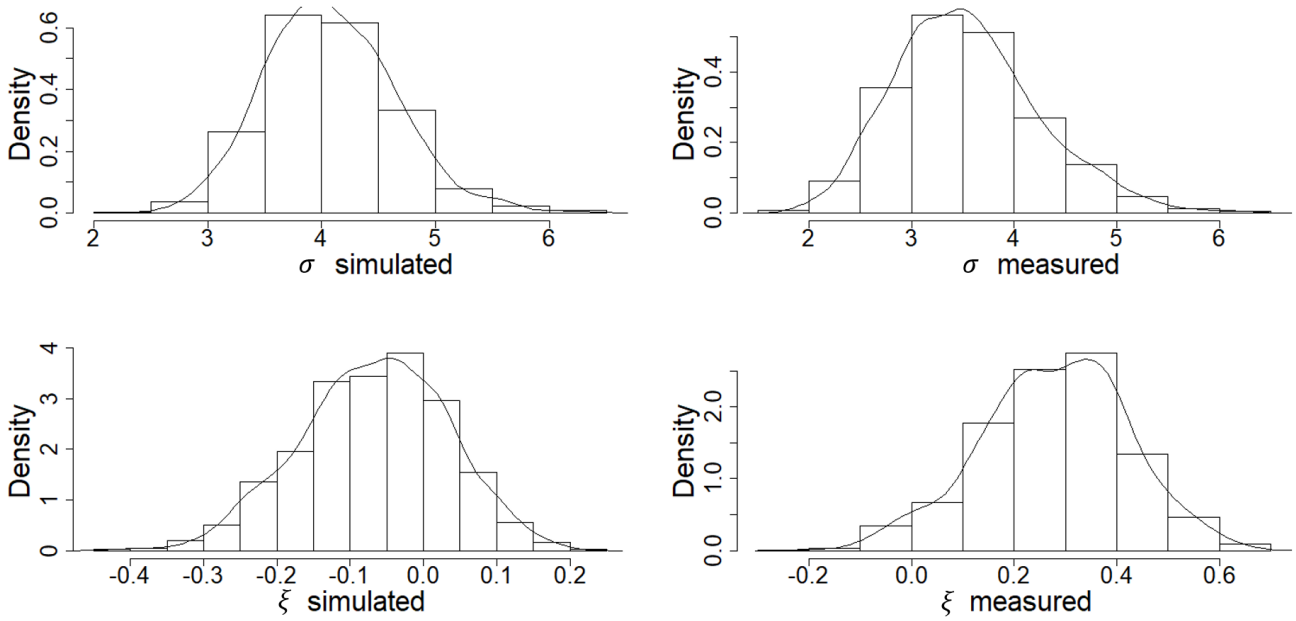


Figure 4-7: Bootstrap-estimated distributions of the scale and shape parameters (top and bottom histograms, respectively) for the conditioning (PBM simulated) and dependent (measured data) variables (left and right histograms, respectively).

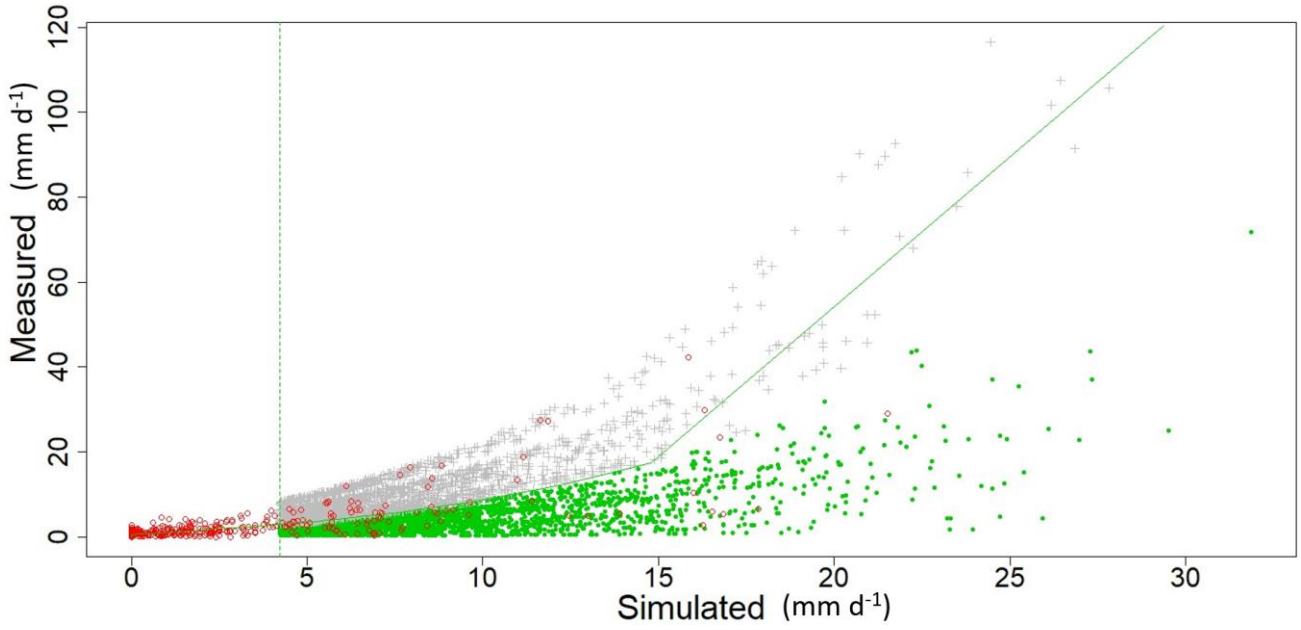


Figure 4-8: Scatterplot of measured versus PBM simulated flow (red circles) together with CEM simulated data (grey crosses and green circles) plotted above the threshold for prediction (green, dashed vertical line). The fitted curve (green solid line) joins equal quantiles of the marginal distributions and is used only for reference.

4.5.4 Hybrid model via CEM-ELM adjustments of PBM simulated data

To recap, this research applies the CEM for the maximum peaks, while the ELM model is used for the smaller peaks during a peak flow event as the ELM alone did not increase the accuracy of the maximum peaks (over that found with the PBM alone). For reference, error and agreement performance indices are given in Appendix C (Figure C-1) for the three constituent models of the study hybrid (i.e. for PBM only, CEM only and ELM only), for predicting the maximum peaks.

The resultant hybrid simulations (or adjusted PBM simulations) for peak flow events above the minimum threshold of 3.88 mm d^{-1} are presented in Figure 4-9 together with the PBM simulated data and the measured data. The PBM most commonly under-predicts the largest peaks and over-predicts the ones preceding and following it. Use of the CEM captures the cluster maxima more accurately, which naturally depends on the value of the PBM simulation. In cases where the PBM over-predicts the maximum peak, the CEM leads to an even greater error. The ELM model addresses the fact that the PBM tends to over-predict the smaller peaks and, thus, provides hybrid forecasts of these peaks that are smaller and closer to the measured ones. The characteristics of the elements of the proposed methodology, in combination, results in improved characterization of the peak flow events, that tend to rise and fall more steeply (and realistically) than is found with the PBM simulations. Key exceptions arise for cases where the PBM over-predicts the whole event, as the hybrid

compounds this over-prediction. Error and agreement indices (Figure 4-10) provide an overall assessment of the proposed hybrid methodology for the same peak flow events (of Figure 4-9), but specifically just for instances of PBM simulations > 3.88 mm d^{-1}. In general, the proposed hybrid approach is more accurate, as it results in smaller error indices and larger agreement indices than produced using the PBM alone, except for PBIAS, despite reductions in the other two error indices (MAE and NRMSE). Clearly, PBIAS is more reflective of how the hybrid can sometimes compound over-prediction. The greatest relative improvement was found in the KGE index, although both NSE and d also indicated improved agreement between observed and hybrid simulated values.

All of the results discussed above relate only to instances of PBM simulated flow values above the threshold of 3.88 mm d^{-1}, where the measured and hybrid simulated values directly correspond to. We compare now between *all* the measured water flow data, the PBM and hybrid simulations when above the selected threshold. The resultant plots of error (MAE and PBIAS only) and agreement (d and KGE only) indices against the magnitude of observed flow are given in Figure 4-11. The MAE is very small for both the PBM and the hybrid when comparing simulated flow with *all* the observed flow above the threshold. Increasing the observed flow threshold above which data are compared with the simulated data, results in a slower increase (with flow magnitude) in the MAE for the hybrid than for the PBM outputs. The hybrid approach also results in a significant decrease of the negative PBIAS with increasing peak flow, relative to the PBM. The agreement indices (d and KGE) similarly confirm this improvement found for the hybrid simulations over the PBM simulations.

All of the results discussed above refer to peak events above the threshold of 3.88 mm d^{-1}, as selected based on the GPD parameter stability plots (Figure 4-5). As a final step in the analysis,

it is prudent to assess how threshold selection has an effect on the performance of the proposed methodology. Thresholds were set to range from 3.88 mm d^{-1} up to the 95th quantile of the PBM simulated flow (6.5 mm d^{-1}). According to the calculated MAE indices, the hybrid model has a performance similar to the PBM when considering peak events above the threshold of 5.8 mm d^{-1} (Figure 4-12). This is not confirmed by the NRMSE which, however, shows a steep increase for the same threshold. PBIAS shows an overall increasing trend with some fluctuations in between. The agreement indices (Figure 4-12) seem to be less sensitive to the threshold, although NSE shows an abrupt decrease when flow is higher than 5.8 mm d^{-1}. All the indices have the common characteristic of the consistent trend (increasing for error, decreasing for agreement) as the threshold increases, which could be attributed to the smaller samples of the data used for testing, in which the highest flow values dominate.

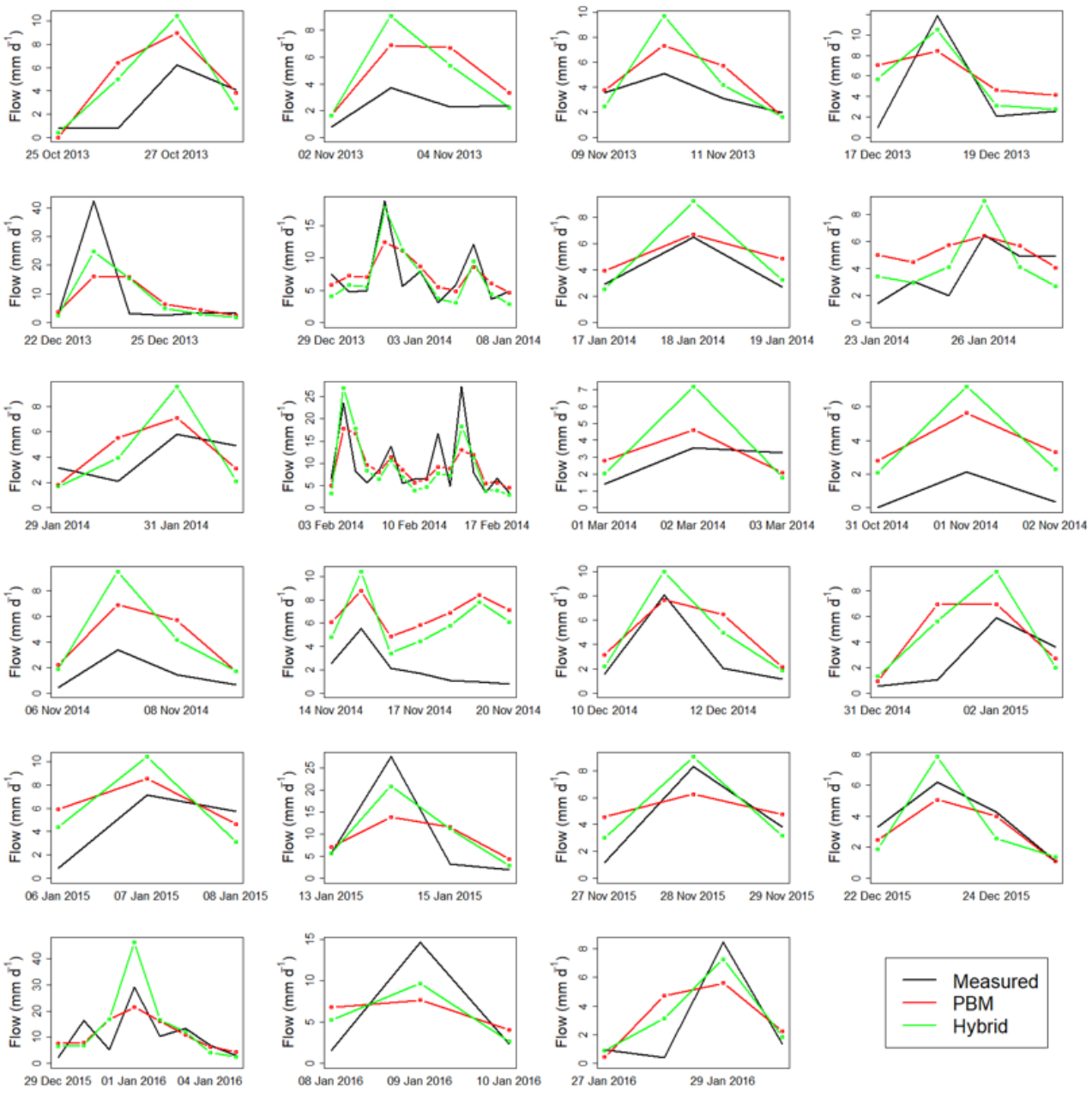


Figure 4-9: Time-series plots of measured, PBM-predicted and hybrid model-predicted flow for all considered peak flow events for which the PBM simulated data > 3.88 mm d^{-1}, following the threshold selection analysis of Section 4.1.

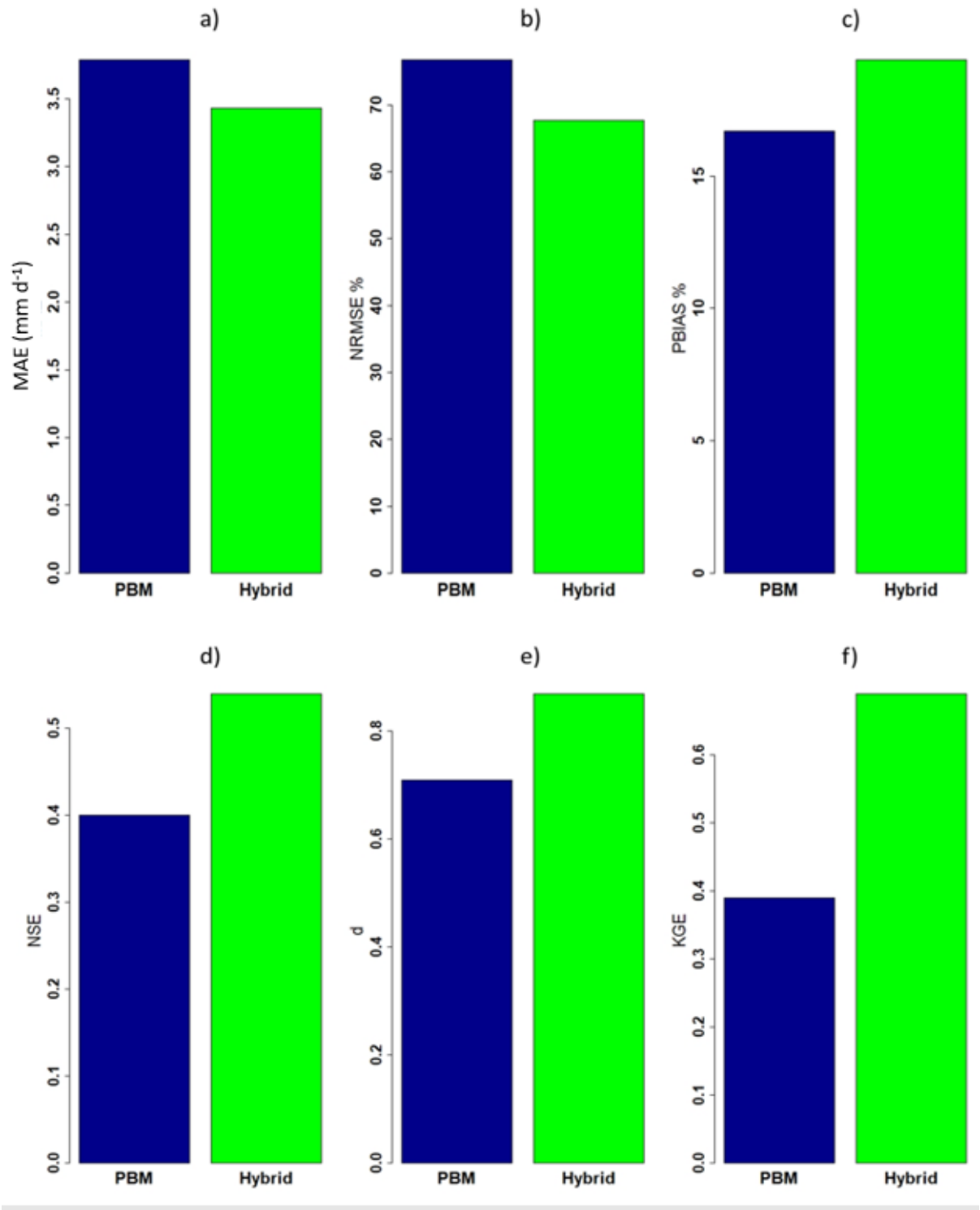


Figure 4-10: Error and agreement indices of the PBM and hybrid simulated peaks compared to observed: a) MAE, b) NRMSE, c) PBIAS, d) NSE, e) d, f) KGE.

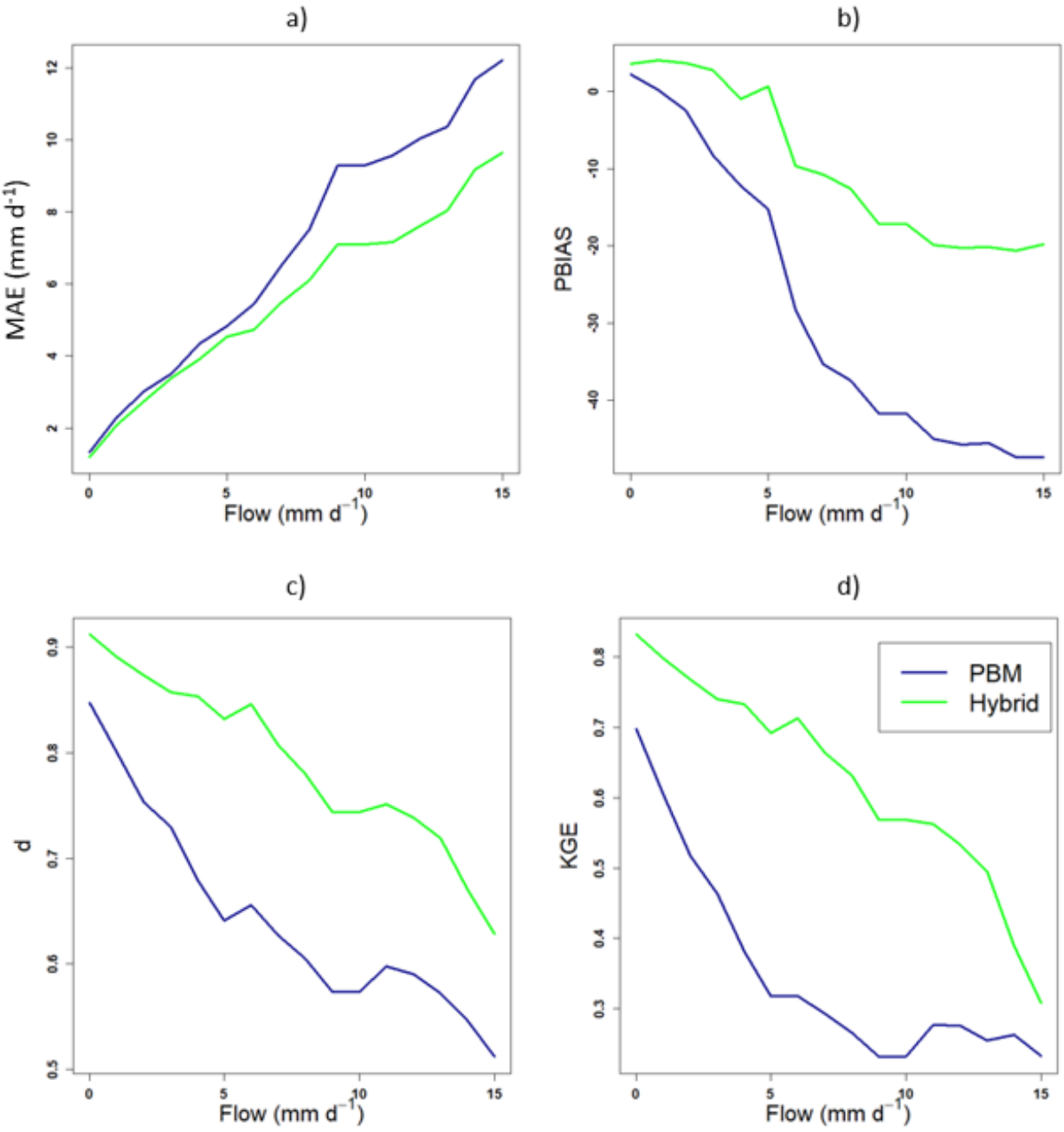


Figure 4-11: Error and agreement indices of the PBM and hybrid simulated data for increasing observed flow values. a) MAE, b) PBIAS, c) d, d) KGE.

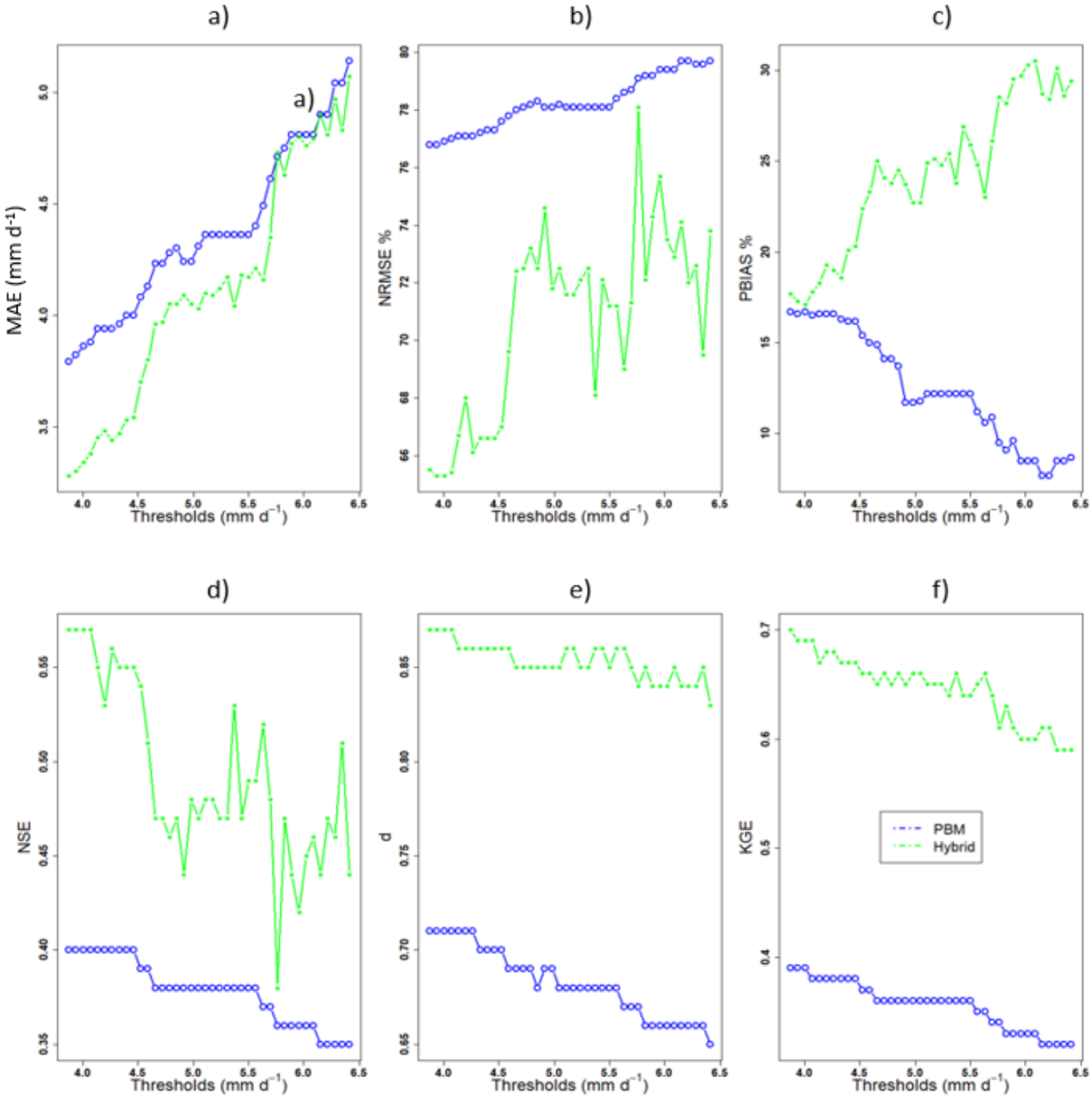


Figure 4-12: Error (MAE, NRMSE, PBIAS; top 3 plots) and agreement (NSE, d, KGE; bottom 3 plots) indices of the PBM and hybrid simulated data for a range of thresholds (3.88 to 6.5 mm d^{-1}).

4.6 Discussion

The main motivation for developing the proposed hybrid approach was to forecast more accurately the peak flows that are typically under-predicted using PBMs due to model over-generalisation or smoothing. The analysis in this research was based on simulations obtained from the SPACSYS model however the hybrid approach presented could be applied to other models that simulate flow. Similar to many other PBMs used for flow simulation, SPACSYS tends to under-predict peak flows and the hybrid approach presented is entirely general. However, the PBM also exhibited other problems, such as over-predicting small and moderate flow values. This second problem arises because the model (as for most PBMs) is calibrated implicitly to the *mean* of the observed distribution through the careful choice and selection of model parameters. It should be noted, however, that SPACSYS is not fitted or re-calibrated explicitly to external data.

Topological characteristics, such as the integrating effect of the catchment, could also contribute to this behaviour. For example, large local slopes (that SPACSYS cannot represent) result in faster running water which, combined with intense rainfall, may result in higher peak flows that are not captured by SPACSYS. Over-predicted events could be due to inaccurate representation of soil moisture, topography and other soil properties at the within-field scale, since SPACSYS simulates at the field scale (Liu et al., 2018). Despite these issues and the fact that our proposed hybrid approach was aimed at under-predicted extreme flow events, the hybrid approach resulted in more accurate forecasts and an increase in accuracy overall.

The CEM is usually used to describe the extreme dependence structure of the same variable at different sites or of different variables at the same site. In this study, we used the CEM in a bivariate context to model and link the same underlying state variable captured by different

representational processes (i.e., direct measurement and PBM simulation of flow). The pseudo-observations obtained from the fitted model and based on the conditioning variable were aggregated to a single value which was then compared to the equivalent measured value. The same conditional simulations can be used to create confidence intervals that correspond to various scenarios and allow flexibility in choosing values according to the intended purpose.

In general, none of the applied criteria for the evaluation of the proposed hybrid method is sufficient singly; each of the model performance indices have strengths and weaknesses. The agreement indices are used mainly to investigate how accurately the model captures the dynamic of the temporal process. The error indices capture differences between the total flow or the volume of the hydrograph. Therefore, using both measures provides a more holistic evaluation of model performance. Since our main objective was to evaluate the performance of the proposed hybrid method in predicting extreme flows, the choice of the agreement indices is appropriate as they have been shown to be sensitive to peaks (Krause et al., 2005).

Despite the promising results obtained from the proposed methodology, it has the limitation of being tested for a specific case study site and for one PBM. Future research should, therefore, consider testing this approach for other catchment sites with different characteristics, as data-driven models need to be tested using a range of (large) datasets before being applied in practice (Boulesteix et al., 2018; Papacharalampous et al., 2019; Tyralis et al., 2019). It would also be interesting to investigate whether and how the performance of SPACSYS, and by extension, the proposed techniques, would be affected by using forecasted weather variables as inputs instead of measured data to obtain the

simulations. In real case scenarios, the threshold is commonly defined based on pre-existing information. Due to the nature of the NWFP experiment, it was not possible to define a threshold with physical meaning (e.g. likely flooding) with which to evaluate the estimated threshold. The threshold defines the peak flow events and consequently the training and testing datasets used in this research. Thus, it was not possible to define a threshold based strictly on the training dataset only as would normally be the case. However, we expect this to have a minimal effect on the results and not change the main conclusions drawn.

4.7 Conclusions

In this research, we used a data-driven machine learning model (ELM) and a semi-parametric conditional model that stems from extreme value theory (CEM) to increase the accuracy of peak water flow events simulated by a process-based model (PBM). The PBM frequently under-predicted the maximum flows during a peak event, for which the CEM was applied, and over-predicted flows preceding and following these peaks, for which the ELM was applied. The combined characteristics of the proposed methodology in general resulted in more accurate forecasts and improved representation of these peak events, according to several error and agreement indices. The detailed analysis undertaken in this research was developed based on simulated flow data obtained from only one PBM and for observed data at only one case study site. However, because of the general characteristics of the chosen PBM and of the proposed hybrid methodology, it is anticipated that the proposed approach will be suitable for a wide range of PBMs and water monitoring station schemes.

Authors contribution statement

Stelian Curceac 80%: Conceptualisation, Methodology, Software, Formal analysis, Writing-Original Draft, Writing-Review & Editing.

Peter Atkinson 5%: Conceptualisation, Writing-Review & Editing, Supervision, Funding acquisition.

Alice Milne 5%: Conceptualisation, Writing-Review & Editing, Supervision, Funding acquisition.

Lianhai Wu 5%: Software, Writing-Review & Editing, Supervision, Funding acquisition.

Paul Harris 5%: Conceptualisation, Data curation, Writing-Review & Editing, Supervision, Funding acquisition.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

Rothamsted Research receives grant aided support from the Biotechnology and Biological Sciences Research Council (BBSRC) of the United Kingdom. This research was funded by Rothamsted Research and Lancaster Environment Centre, the BBSRC Institute Strategic Programme (ISP) grant, "Soils to Nutrition" (S2N) grant numbers BBS/E/C/000I0320, BBS/E/C/000I0330 and the BBSRC National Capability grant for the North Wyke Farm Platform grant number BBS/E/C/000J0100. The authors wish to thank the Editor and two anonymous reviewers for their useful comments, which led to considerable improvements to the paper.

Data and Software Availability Statement

All North Wyke Farm Platform datasets (https://www.rothamsted.ac.uk/north-wyke-farm-platform) and the SPACSYS model (https://www.rothamsted.ac.uk/rothamsted-spacsys-model) are freely available. R software (R Core Team, 2019) was used for the implementation

of the statistical models. The CEM was applied by using the texmex R package (Southworth et al., 2018), the elmNNRcpp R package was used for the ELM model (Mouselimis and Gosso, 2018) and the indices were calculated by using functions in the hydroGOF R package (Zambrano-Bigiarini, 2017).

References

Bates, B. C., Kundzewicz, Z. W., Wu, S. and Palutikof, J. P. (2008). Climate Change and Water. Technical Paper of the Intergovernmental Panel on Climate Change, IPCC Secretariat, Geneva, 210 pp.

Bogner, K., Liechti, K. and Zappa, M. (2016). Post-Processing of Stream Flows in Switzerland with an Emphasis on Low Flows and Floods, Water, 8 (4), 115. doi: 10.3390/w8040115.

Bogner, K., Liechti, K. and Zappa, M. (2017). Technical Note: Combining Quantile Forecasts and Predictive Distributions of Streamflows, Hydrology and Earth System Sciences, 21 (11), 5493-5502. doi: 10.5194/hess-21-5493-2017.

Boulesteix, A. L., Binder, H., Abrahamowicz, M. and Sauerbrei, W. (2018). On the Necessity and Design of Studies Comparing Statistical Methods, Biometrical Journal, 60 (1), 216-218. doi: 10.1002/bimj.201700129.

Bouraoui, F., Grizzetti, B., Granlund, K., Rekolainen, S. and Bidoglio, G. (2004). Impact of Climate Change on the Water Cycle and Nutrient Losses in a Finnish Catchment, Climatic Change, 66(1–2), 109-126. doi.org: 10.1023/B:CLIM.0000043147.09365.e3.

Box, G. E. P. and Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control, Revised Edition. Holden-Day, San Francisco, CA.

Bradley, A. A., Habib, M. and Schwartz, S. S. (2015). Climate index weighting of ensemble streamflow forecasts using a simple Bayesian approach, Water Resources Research, 51, 7382–7400. doi: 10.1002/2014WR016811.

Chen, L., Sun, N., Zhou, C., Zhou, J., Zhou, Y., Zhang, J. and Zhou, Q . (2018). Flood Forecasting Based on an Improved Extreme Learning Machine Model Combined with the Backtracking Search Optimization Algorithm, Water, 10(10), 1362. doi: 10.3390/w10101362.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots, Journal of the American Statistical Association, 74(368), 829-836. doi: 10.2307/2286407.

Cloke, H. L. and Pappenberger, F. (2009). Ensemble Flood Forecasting: A Review, Journal of Hydrology, 375(3), 613-626. doi: 10.1016/j.jhydrol.2009.06.005.

Collet, L., Beevers, L. and Prudhomme, C. (2017). Assessing the Impact of Climate Change and Extreme Value Uncertainty to Extreme Flows across Great Britain, Water, 9(2), 103. doi: 10.3390/w9020103.

Curceac, S., Atkinson, P. M., Milne, A., Wu, L. and Harris, P. (2020). An Evaluation of Automated GPD Threshold Selection Methods for Hydrological Extremes across Different Scales, Journal of Hydrology, 585, 124845. doi: 10.1016/j.jhydrol.2020.124845.

Deo, R. C. and Şahin, M. (2016). An Extreme Learning Machine Model for the Simulation of Monthly Mean Streamflow Water Level in Eastern Queensland, Environmental Monitoring and Assessment, 188, 90. doi: 10.1007/s10661-016-5094-9.

Dogulu, N., López López, P., Solomatine, D. P., Weerts, A. H. and Shrestha, D. L. (2015). Estimation of Predictive Hydrologic Uncertainty Using the Quantile Regression and UNEEC

Methods and Their Comparison on Contrasting Catchments, Hydrology and Earth System Sciences, 19 (7), 3181-3201. doi: 10.5194/hess-19-3181-2015.

Drees, H. and Janßen, A. (2017). Conditional Extreme Value Models: Fallacies and Pitfalls, Extremes, 20(4), 777–805. doi: 10.1007/s10687-017-0293-5.

Fathian, F., Mehdizadeh, S., Kozekalani A. S. and Safari, M. J. S. (2019). Hybrid Models to Improve the Monthly River Flow Prediction: Integrating Artificial Intelligence and Non-Linear Time Series Models, Journal of Hydrology, 575, 1200–1213. doi: 10.1016/j.jhydrol.2019.06.025.

Field, C. B., Barros, V., Stocker, T. F. and Dahe, Q. (2012). Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change, Cambridge, Cambridge University Press. doi: 10.1017/CBO9781139177245.

Heffernan, J. E. and Tawn, J. A. (2004). A Conditional Approach for Multivariate Extreme Values (with Discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(3), 497-546. doi.org: 10.1111/j.1467-9868.2004.02050.x.

Huang, G. B., Zhu, Q. Y. and Siew, C. K. (2006). Extreme Learning Machine: Theory and Applications, Neurocomputing, Neural Networks, 70(1), 489-501. doi: 10.1016/j.neucom.2005.12.126.

Keef, C., Papastathopoulos, I. and Tawn, J. A. (2013). Estimation of the Conditional Distribution of a Multivariate Variable given That One of Its Components Is Large: Additional Constraints for the Heffernan and Tawn Model, Journal of Multivariate Analysis, 115, 396-404. doi: 10.1016/j.jmva.2012.10.012.

Kisi, O. and Cimen, M. (2011). A Wavelet-Support Vector Machine Conjunction Model for Monthly Streamflow Forecasting, Journal of Hydrology, 399(1), 132-140. doi: 10.1016/j.jhydrol.2010.12.041.

Krause, P., Boyle, D. P. and Bäse, F. (2005). Comparison of Different Efficiency Criteria for Hydrological Model Assessment, Advances in Geosciences, 5, 89-97. doi: 10.5194/adgeo-5-89-2005.

Kundzewicz, Z. W., Mata, L. J., Arnell, N. W., Doll, P., Kabat, P., Jimenez, B. et al. (2007). Freshwater Resources and Their Management. In Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by M. L. Parry, O. F. Canziani, J. P. Palutikof, P. J. van der Linden, and C. E. Hanson, 173–210. Cambridge University Press.

Lamb, R., Keef, C., Tawn, J., Laeger, S., Meadowcroft, I., Surendran, S., Dunning, P. and Batstone, C. (2010). A New Method to Assess the Risk of Local and Widespread Flooding on Rivers and Coasts, Journal of Flood Risk Management, 3(4), 323-336. doi: 10.1111/j.1753-318X.2010.01081.x.

Lane, R. A., Coxon, G., Freer, J. E., Wagener, T., Johnes, P. J., Bloomfield, J. P., Greene, S., Macleod, C. J. A. and Reaney, S. M. (2019). Benchmarking the Predictive Capability of Hydrological Models for River Flow and Flood Peak Predictions across over 1000 Catchments in Great Britain, Hydrology and Earth System Sciences, 23(10), 4011-4032. doi: 10.5194/hess-23-4011-2019.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W. and Di, Z. (2017). A Review on Statistical Postprocessing Methods for Hydrometeorological Ensemble Forecasting. Wiley Interdisciplinary Reviews, Water, 4(6): e1246. doi: 10.1002/wat2.1246.

Li, X. Q., Chen, J., Xu, C. Y., Li, L. and Chen, H. (2019). Performance of Post-Processed Methods in Hydrological Predictions Evaluated by Deterministic and Probabilistic Criteria, Water Resources Management, 33(9), 3289-3302. doi: 10.1007/s11269-019-02302-y.

Lima, A. R., Cannon, A. J. and Hsieh, W. W. (2015). Nonlinear Regression in Environmental Sciences Using Extreme Learning Machines: A Comparative Evaluation, Environmental Modelling & Software, 73, 175-188. doi: 10.1016/j.envsoft.2015.08.002.

Liu, Y., Li, Y., Harris, P., Cardenas, L. M., Dunn, R. M., Sint, H., Murray, P. J., Lee, M. R. F. and Wu, L. (2018). Modelling Field Scale Spatial Variation in Water Run-off, Soil Moisture, N2O Emissions and Herbage Biomass of a Grazed Pasture Using the SPACSYS Model, Geoderma, 315, 49-58. doi: 10.1016/j.geoderma.2017.11.029.

López López, P., Verkade, J. S., Weerts, A. H. and Solomatine, D. P. (2014). Alternative Configurations of Quantile Regression for Estimating Predictive Uncertainty in Water Level Forecasts for the Upper Severn River: A Comparison. Hydrology and Earth System Sciences, 18(9), 3411-3428. doi: 10.5194/hess-18-3411-2014.

McCuen R. H. (2005). Accuracy Assessment of Peak Discharge Models, Journal of Hydrologic Engineering, 10, (1), 16-22. doi: 10.1061/(ASCE)1084-0699(2005)10:1(16).

Mendes, B. V. de M. and Pericchi, L. R. (2009). Assessing Conditional Extremal Risk of Flooding in Puerto Rico, Stochastic Environmental Research and Risk Assessment, 23(3), 399-410. doi: 10.1007/s00477-008-0220-z.

Miller, R. G. (1964). A Trustworthy Jackknife, The Annals of Mathematical Statistics, 35(4), 1594-1605. doi: 10.1214/aoms/1177700384.

Mouselimis, L. and Gosso, A. (2018). elmNNRcpp: The Extreme Learning Machine Algorithm. R package version 1.0.1. https://CRAN.R-project.org/package=elmNNRcpp

Nash, J. E. and Sutcliffe, J. V. (1970). River Flow Forecasting through Conceptual Models Part I - A Discussion of Principles, Journal of Hydrology, 10, (3): 282-290. doi: 10.1016/0022-1694(70)90255-6.

Orr, R. J., Murray, P. J., Eyles, C. J., Blackwell, M. S. A., Cardenas, L. M., Collins, A. L. et al. (2016). The North Wyke Farm Platform: effect of temperate grassland farming systems on soil moisture contents, runoff and associated water quality dynamics, European Journal of Soil Science, 67, 374–385. doi: 10.1111/ejss.12350.

Papacharalampous, G., Tyralis, H., Langousis, A., Jayawardena, A. W., Sivakumar, B., Mamassis, N., Montanari, A. and Koutsoyiannis, D. (2019). Probabilistic Hydrological Post-Processing at Scale: Why and How to Apply Machine-Learning Quantile Regression Algorithms, Water, 11(10), 2126. doi: 10.3390/w11102126.

Quilty, J., Adamowski, J. and Boucher, M. A. (2019). A Stochastic Data-Driven Ensemble Forecasting Framework for Water Resources: A Case Study Using Ensemble Members Derived From a Database of Deterministic Wavelet-Based Models, Water Resources Research, 55(1), 175-202. doi: 10.1029/2018WR023205.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Monthly Weather Review, 133(5), 1155-1574. doi: 10.1175/MWR2906.1.

Roulin, E. and Vannitsem, S. (2011). Postprocessing of Ensemble Precipitation Predictions with Extended Logistic Regression Based on Hindcasts, Monthly Weather Review, 140(3), 874-888. doi: 10.1175/MWR-D-11-00062.1.

Scarrott, C. and MacDonald, A. (2012). A Review of Extreme Value Threshold Es-Timation and Uncertainty Quantification, REVSTAT–Statistical Journal, 10(1), 33-60.

Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap, Springer Series in Statistics, New York, Springer-Verlag. doi: 10.1007/978-1-4612-0795-5.

Sikorska, A. E., Montanari, A. and Koutsoyiannis, D. (2015). Estimating the Uncertainty of Hydrological Predictions through Data-Driven Resampling Techniques, Journal of Hydrologic Engineering, 20(1), A4014009. doi: 10.1061/(ASCE)HE.1943-5584.0000926.

Southworth, H., Heffernan J. E. and Metcalfe, P. D. (2018). texmex: Statistical modelling of extreme values. R package version 2.4.2.

Sun, Z. L., Choi, T. M., Au, K. F. and Yu, Y. (2008). Sales Forecasting Using Extreme Learning Machine with Applications in Fashion Retailing, Decision Support Systems, 46(1), 411-419. doi: 10.1016/j.dss.2008.07.009.

Takahashi, T., Harris, P. M., Blackwell, S. A., Cardenas, L. M., Collins, A. L., Dungait, J. A. J., Hawkins, J. M. B. et al. (2018). Roles of Instrumented Farm-Scale Trials in Trade-off Assessments of Pasture-Based Ruminant Production Systems, Animal, 12(8),1766-1776. doi: 10.1017/S1751731118000502.

Thibault, K. M. and Brown, J. H. (2008). Impact of an Extreme Climatic Event on Community Assembly, Proceedings of the National Academy of Sciences, 105(9), 3410-3415. doi: 10.1073/pnas.0712282105.

Toth, E., Montanari, A. and Brath, A. (1999). Real-Time Flood Forecasting via Combined Use of Conceptual and Stochastic Models, Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere, 24(7), 793-798. doi: 10.1016/S1464-1909(99)00082-9.

Tyralis, H., Papacharalampous, G., Burnetas, A. and Langousis, A. (2019). Hydrological Post-Processing Using Stacked Generalization of Quantile Regression Algorithms: Large-Scale Application over CONUS, Journal of Hydrology, 577, 123957. doi: 10.1016/j.jhydrol.2019.123957.

Wijayarathne, D. B. and Coulibaly, P. (2020). Identification of Hydrological Models for Operational Flood Forecasting in St. John's, Newfoundland, Canada, Journal of Hydrology: Regional Studies, 27, 100646. doi: 10.1016/j.ejrh.2019.100646.

Wu, L., McGechan, M. B. McRoberts, N., Baddeley, J. A. and Watson, C. A. (2007). SPACSYS: Integration of a 3D Root Architecture Component to Carbon, Nitrogen and Water Cycling-Model Description, Ecological Modelling, 200(3), 343-359. doi: 10.1016/j.ecolmodel.2006.08.010.

Yaseen, Z. M. Jaafar, O., Deo, R. C., Kisi, O., Adamowski, J., Quilty, J. and El-Shafie, A. (2016). Stream-Flow Forecasting Using Extreme Learning Machines: A Case Study in a Semi-Arid Region in Iraq, Journal of Hydrology, 542, 603-614. doi: 10.1016/j.jhydrol.2016.09.035.

Yaseen, Z. M., Sulaiman, S. O., Deo, R. C. and Chau, K. W. (2019). An Enhanced Extreme Learning Machine Model for River Flow Forecasting: State-of-the-Art, Practical Applications in Water Resource Engineering Area and Future Research Direction, Journal of Hydrology, 569, 387-408. doi: 10.1016/j.jhydrol.2018.11.069.

Zambrano-Bigiarini, M. (2017). hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time seriesR package version 0.3-10.http://hzambran.github.io/hydroGOF/. DOI:10.5281/zenodo.840087.

Zheng, F., Westra, S., Leonard, M. and Sisson, S. A. (2014). Modelling Dependence between Extreme Rainfall and Storm Surge to Estimate Coastal Flooding Risk, Water Resources Research, 50(3), 2050-2071. doi: 10.1002/2013WR014616.

Zhou, J., Peng, T., Zhang, C. and Sun, N. (2018). Data Pre-Analysis and Ensemble of Various Artificial Neural Networks for Monthly Streamflow Forecasting, Water, 10(5), 628. doi: 10.3390/w10050628.

5. Elucidating the performance of hybrid models for predicting extreme water flow events through variography and wavelet analyses

Stelian Curceac[a*], Alice Milne[b], Peter M. Atkinson[c,d,e], Lianhai Wu[a], Paul Harris[a]

[a] Rothamsted Research, Department of Sustainable Agriculture Sciences, North Wyke EX20 2SB, Devon, UK.

[b]Rothamsted Research, Department of Sustainable Agriculture Sciences, Harpenden AL5 2JQ, UK

[c]Lancaster Environment Centre, Lancaster University, Bailrigg, Lancaster LA1 4YQ, UK.

[d]Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

[e]State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China.

***Correspondence:**

Stelian Curceac

stelian.curceac@rothamsted.ac.uk

Published in Journal of Hydrology

5.1 Abstract

Accurate prediction of extreme flow events is important for mitigating natural disasters such as flooding. We explore and refine two modelling approaches (both separately and in combination) that have been demonstrated to improve the prediction of daily peak flow events. These are hybrid models that combine process-based models (PBM) with statistical and machine learning methods and models that aggregate fine resolution (sub-daily) PBM simulated flow to daily. We propose the use of variography and wavelet analyses to evaluate these models across temporal scales. These exploratory methods are applied to both measured and modelled data in order to assess the performance of the latter in capturing variation, at different scales, of the former. Critically, we compare change points detected by the wavelet analysis (measured and modelled) with the extreme flow events identified in the measured data. We found that combining the two modelling approaches improves prediction at finer scales but at coarser scales advantages are less pronounced. Although aggregating fine-scale model outputs improved the partition of wavelet variation across scales, the autocorrelation in the signal is less well represented as demonstrated by variography. We demonstrate that exploratory time-series analyses, using variograms and wavelets, provides a useful assessment of existing and newly proposed models, with respect to how they capture changes in flow variance at different scales and also how this correlates with measured flow data – all in the context of extreme flow events.

Keywords

Variogram analysis, wavelet analysis, process scale, peak flows, hydrology

5.2 1. Introduction

In many regions across the globe, changing patterns of rainfall have increased the risk of extreme water flows and associated flooding, posing unique challenges for both urban and rural environments (Bates et al., 2008; Field et al., 2012; Kundzewicz et al., 2007). Whereas in urban environments, homes and businesses may be at risk of severe damage, in rural environments, agricultural production can be at risk through waterlogging (Brown et al., 2016), soils may be threatened by erosion and watercourses may become contaminated by excess nutrients as a result of fertilizer in runoff (Bouraoui et al., 2004). To manage and mitigate the impacts of extreme flow events accurate and reliable modelling and forecasting of flow, and particularly extreme flow events, are needed.

Catchment hydrology has been modelled using mechanistic or semi-empirical models (e.g. Jaiswal et al., 2020), in which known processes are described. These models tend to capture the coarse scale variation in observed flow relatively well. However, fine-scale variation is often under-predicted reducing the accuracy of forecasting the true magnitude of extreme events. Wu et al. (2020) investigated the effect of the simulation time-step on predicting extreme daily flows (mm day^{-1}) and discovered that using finer resolution input data and then aggregating the process-based model (PBM) outputs to the daily scale increased accuracy, both in the prediction of general trends and identification of peak flows. In effect, the hydrological model functions as a filter or transform which reduces the influence of high frequency weather variation. When input data are aggregated (e.g., from hourly to daily resolution) the variation is damped through averaging over this larger 'support', which is appropriate. However, the model filter may dampen the variation still further resulting in under-prediction of extreme events. Aggregation of model outputs generated from fine-

resolution inputs tends to retain better the extreme peaks in the data because the dampening effect of the model is restricted to the hourly time-step.

An increasingly popular approach to increase the accuracy of the prediction of extreme events is to fit hybrid models (e.g. Curceac et al., 2020a). These models integrate PBM outputs and statistical data-driven methods such as those based on machine learning. For example, in the case of Curceac et al. (2020a), a conditional extreme model (CEM) (Heffernan and Tawn, 2004) and an extreme learning machine (ELM) (Huang et al., 2006) were used to increase the accuracy of simulations of peak flow events obtained from the PBM. An essential element of a hybrid formulation is the ability of the PBM to predict the timing of extreme events.

Key to the accuracy of model predictions of fine-scale extreme events is how well the model captures the underlying processes across scales. Here, we propose the use of variograms and wavelet analysis as tools to explore and assess model performance in characterising temporal patterns in the data across scales. The variogram is the principal tool of geostatistics and, as such, has been used to describe complex variation in spatial data (Goovaerts, 1997; Chilès and Delfiner, 2009; Gringarten and Deutsch, 2001; San Martín et al., 2018). A variogram provides a global (stationary) assessment of spatial or temporal dependence or autocorrelation. For temporal applications, it is able to identify the temporal scales over which a stochastic process is autocorrelated, as well identify any periodicities in the data. Whereas variograms provide a global assessment of temporal dependence in time-series data, wavelet analyses provides a local (non-stationary) assessment across various scales or decompositions (Percival and Guttorp, 1994; Lark and Webster, 1999; Percival and Walden, 2000; Rust et al., 2014). Transforming a time-series by wavelets results in a set of wavelet coefficients, each of which describes the local variation of the signal within a certain scale interval. These coefficients can

be used to determine how the variance (or correlation in the case of two time-series) is partitioned across scales. Changes in the variance of the time-series for a particular scale interval is reflected in the wavelet coefficients and, as such, it is also possible to detect significant changes in the variation for a given scale interval over time.

In this research, the modelling concepts described above are integrated to explore the relative increases in accuracy possible by aggregating fine-scale model outputs and hybrid models, and a combination of the two. Specifically, hybrid models are formed using both the direct daily simulations of a conventional PBM and the aggregation-based PBM outputs. Further, we explore using measured soil moisture as a covariate in the ELM part of the hybrid models. Variograms are used to investigate the existence of nested scales of variation in the measured flow data and assess how (or if) this is captured in the modelled flow data. Wavelet analyses are similarly applied to both measured and modelled flow data to assess the performance of the latter in capturing variation of the former at different scales and locations in time. Critically, we compare change points detected by the wavelet analysis (measured and modelled) with the extreme flow events suggested by the threshold selected based on stability plots of the Generalized Pareto distribution (GPD) of (Curceac et al., 2020b). The exploratory analyses using variograms and wavelets presented here provides a useful assessment of existing and newly proposed models, with respect to how they capture changes in variance at different scales and also how this correlates with measured data; all in the context of extreme flow events. The approach extends that given in Rust et al. (2014), where measured and modelled data were compared using wavelets with respect to changes in land use and management. The approach provides complementary model assessments to those undertaken more routinely based on model prediction accuracy through accuracy metrics

such as are produced by, for example, cross-validation (Smith et al., 1997). Taken together, increased understanding of peak flow processes together with increased peak flow detection accuracy has the potential to provide clear management benefits, not only in flood forecasting, but also reducing nutrient losses to water in an agricultural context.

5.3 Materials and Methods

5.3.1 Study site and data

Water flow data were measured at the North Wyke Farm Platform (NWFP), SW England (50°46'10"N, 3°54'05"W). The NWFP is a farm-scale experiment that was established in 2010 to facilitate research into sustainable grassland livestock systems (Orr et al., 2016; Takahashi et al., 2018). For the period 1985-2015, the mean annual temperature at North Wyke ranges from 6.8 to 13.4 °C and the mean annual rainfall is 1033 mm. The platform's altitude ranges from 120–180 m above sea level. Soil texture consists of a slightly stony clay loam topsoil (approximately 36% clay) above a mottled stony clay (approximately 60% clay). The subsoil is impermeable to water and during rain events most of the excess water moves by surface and subsurface lateral flow towards the drainage system described below. The platform comprises 15 sub-catchments (inset in Figure 5-1) all of which are hydrologically isolated through a combination of topography and a network of French drains (800 mm deep trenches). This ensures that the total runoff is channelled to instrumented flumes, measuring water discharge and water chemistry. For all sub-catchments, runoff has been measured at a 15-minute temporal frequency since October 2012 through a combination of primary and secondary flow devices (as detailed in Orr et al., 2016; Curceac et al., 2020a). The flow is generated only from rainfall as the fields are not irrigated. Further details on the NWFP are given in Section 1.7.

Each sub-catchment has a soil moisture station (SMS) sited at a central location (Figure 1-2), consisting of a remote telemetry unit (RTU), a combination of soil moisture (SM) and temperature probe and a rain gauge (RG) [Adcon, Austria]. The SM probe measures SM through capacitance at depths of 10cm, 20cm and 30cm, and soil temperature at 15cm. However, only SM data at 10cm are available on the data portal as data at the lower depths were deemed unreliable for this soil series. The direct connection to the RTU is via a SDI 12 interface and the raw data is converted to SM using a lookup table developed from testing the sensor output in blocks of North Wyke soil at a range of conditions. For this research, we used discharge, rainfall and SM (from April 2013 to February 2016) measured at sub-catchment 6 (Figure 5-1 and Figure 5-2), which consists of a single field (Golden Rove). This field was chosen because, as part of the permanent pasture treatment of the NWFP, it would not have been ploughed and reseeded during the period of study (which would affect the run-off process).

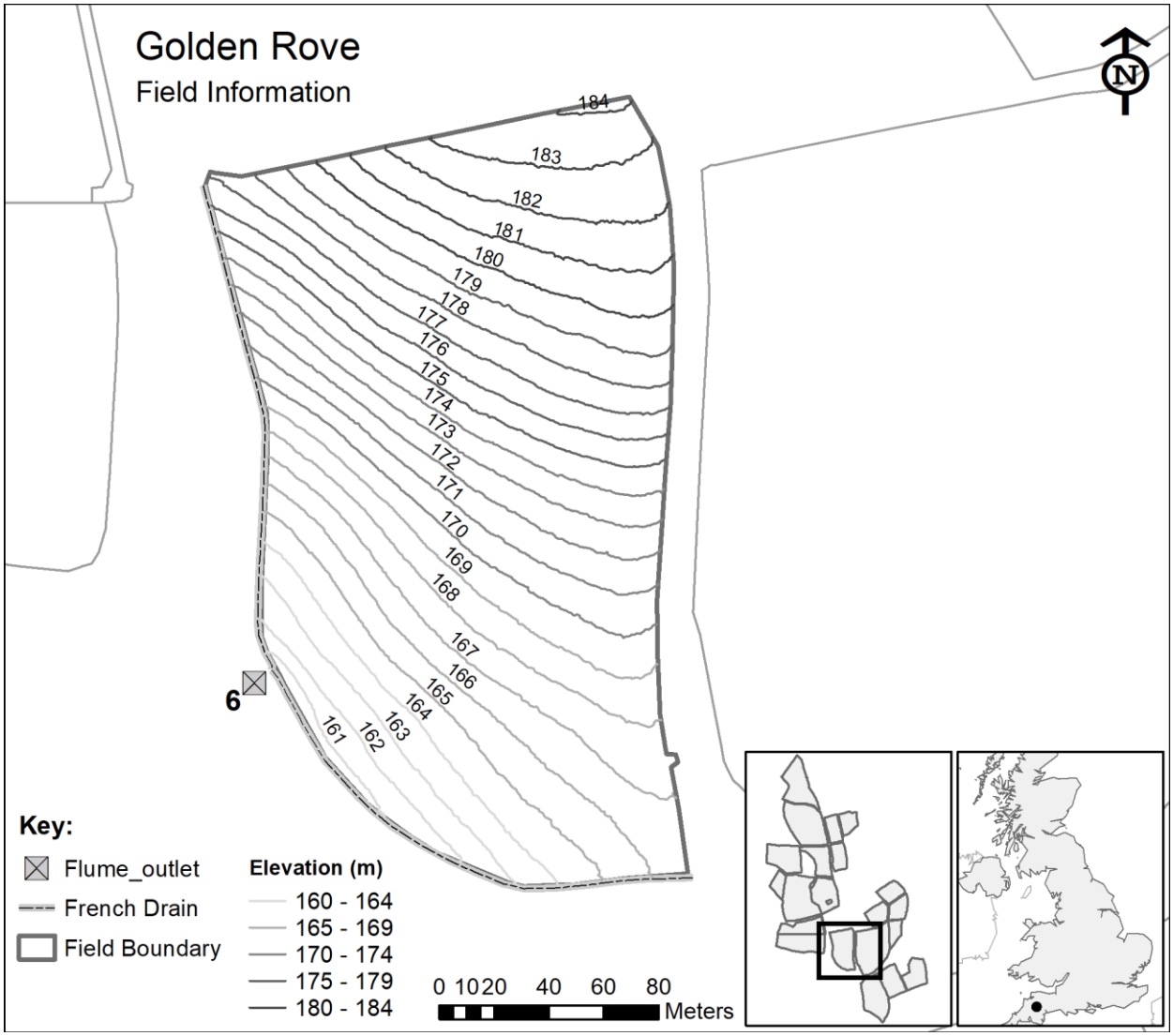


Figure 5-1: Sub-catchment (consisting of a single field) selected from the total of 15 sub-catchments within the North Wyke Farm Platform, South-West England, UK. Precipitation and soil moisture data are collected from a rain gauge and soil moisture site centrally-located in the sub-catchment (see Section 1.7 and Figure 1-2).

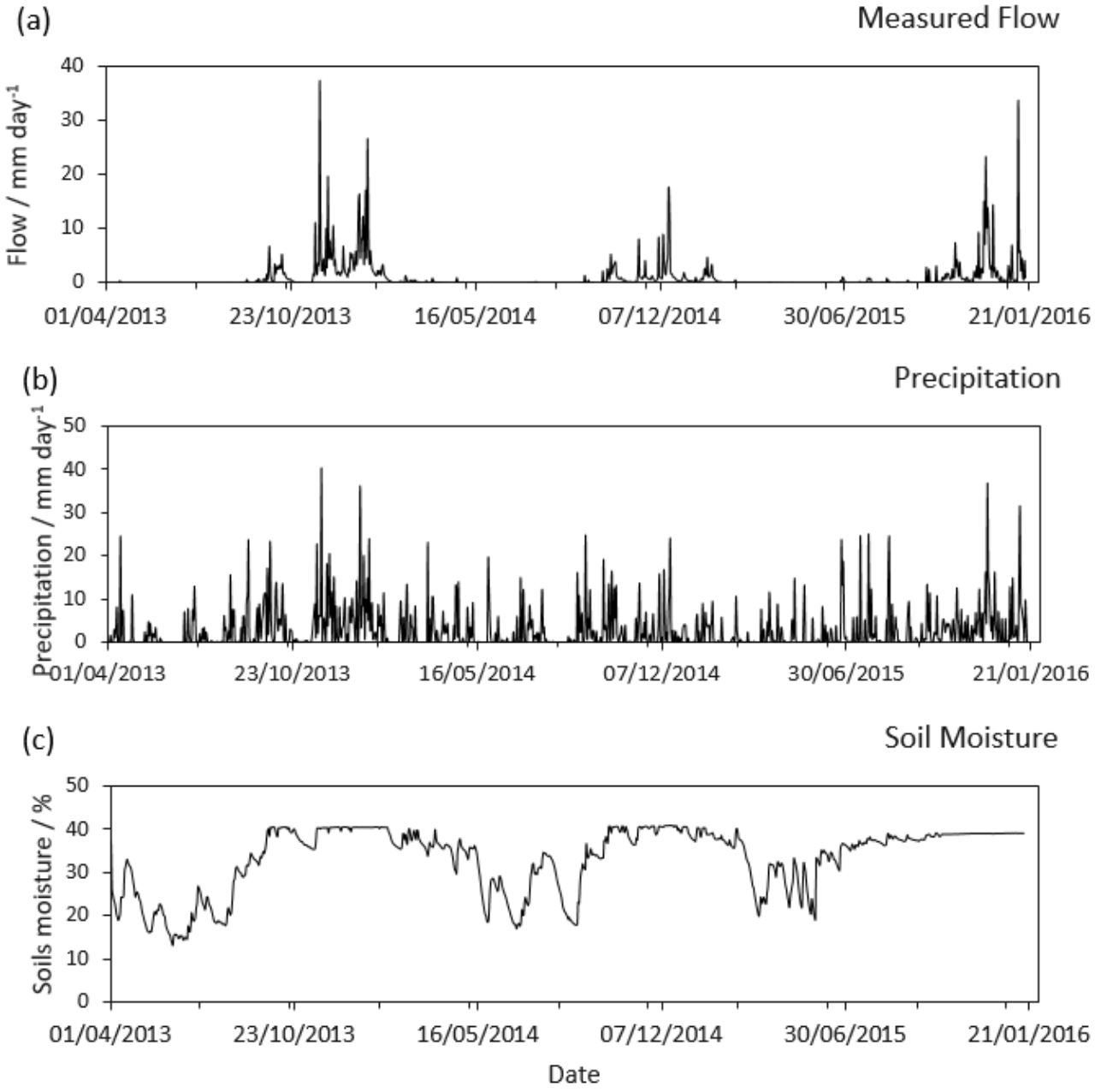


Figure 5-2: (a) Flow data (mm d^{-1}) measured at the study site, (b) precipitation (mm d^{-1}) used as input in the PBM and (c) soil moisture (volumetric %) used as a covariate in the ELM component of the hybrid model. All measurements aggregated from 15 minute to daily.

5.3.2 Models for simulation and forecasting

5.3.2.1 Process-based Model (PBM)

Discharges for the sub-catchment over the period of interest were simulated using the 'SPACSYS' model. SPACSYS is a process-based, field-scale model which simulates key agricultural processes such as plant growth and development, soil carbon and nitrogen cycling, water dynamics and heat transformation (Wu et al., 2007). Water redistribution in a soil profile is simulated by the Richards' equation. Site-specific input data include weather variables (i.e. rainfall) at a given time-step, soil properties, and crop and field management (e.g., fertiliser application rates, composition and dates, grazing and cutting dates). A detailed explanation of SPACSYS including previous simulations of water run-off, soil moisture and other agricultural processes for the same sub-catchment of the NWFP can be found in Liu et al. (2018), where a detailed explanation on the SPACSYS calibration is given.

5.3.2.2 Daily and hourly-to-daily simulations using PBM

SPACSYS has been parameterised to run at 15-minute, hourly, 6 hourly and daily time-steps, depending on the input weather variables available to run the simulations (Wu et al., 2020). For this research, we used simulated discharges at both daily resolution and hourly resolution aggregated to a daily form. The reason for the latter approach was that it was found to increase accuracy in predicting general trends and the identification of peak flows compared to the simulations applied on a daily time-step (Wu et al., 2020).

5.3.2.3 Hybrid PBM with statistical and machine learning models

Following Curceac et al. (2020a), the simulated peak flows obtained from the PBM were post-processed using two approaches, namely the CEM and ELM models.

5.3.2.3.1 GPD and threshold selection

Initially, the extreme flows were fitted by the Generalised Pareto distribution (GPD), with a cumulative distribution function (CDF):

$$G(x) = \Pr(X - u < x | X > u) = \begin{cases} 1 - \left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-\frac{1}{\xi}}, \xi \neq 0 \\ 1 - e^{\left(-\frac{x-u}{\sigma}\right)}, \quad \xi = 0 \end{cases} \qquad 5\text{-}1$$

where x, for this research is the peak flow in mm d^{-1}, u is the location parameter, σ the scale parameter and ξ the shape parameter. The location parameter is the threshold above which flows are considered extreme. A high enough threshold reduces the bias as the GPD is a satisfactory fit to the tail of the empirical distribution, but results in a small sample size which increases the variance. A threshold that is too low results in a large sample size but increases the bias as the empirical distribution deviates from the perfect GPD. According to Extreme Value Theory, if the GPD is a suitable model for the excesses above a high enough threshold u, then it will also be appropriate for all higher thresholds u^* with the shape ξ and modified scale $\sigma_1 = \sigma_{u^*} - \xi u$ being relatively constant (Coles, 2001; Scarrott and MacDonald, 2012). As in Curceac et al. (2020b), we fitted cubic splines to the estimated shape and modified scale parameters for a range of thresholds and calculated the minimum change range which locates the most stable part.

5.3.2.3.2 CEM

For a continuous d-dimensional vector variable $X = (X_1, \dots, X_d)$ with unknown distribution function $F(x)$, the CEM describes the conditional distribution of $X_{-i}|X_i > u_{X_i}$, where X_{-i} is the vector variable X excluding the component X_i. The marginal distribution of each $X_i, i = 1, \dots, d$ is estimated by the GPD model as described above. This can provide different

distributions depending on the shape parameters of the GPD. Therefore, all the components are transformed to either the Gumbel or Laplace distribution for them to follow the same margins. The initial vector variable X is, therefore, transformed as:

$$f(x) = \begin{cases} \log\{2F_{X_i}(X_i)\}, & X_i < F_{X_i}^{-1}(0.5) \\ -\log\{2[1 - 2F_{X_i}(X_i)]\}, & X_i \geq F_{X_i}^{-1}(0.5) \end{cases} \quad 5\text{-}2$$

where $F_{X_i}^{-1}$ is the inverse cumulative distribution function of X_i. The resulting vector variable $Y = (Y_1, \dots, Y_d)$, therefore, has Laplace margins with:

$$\Pr(Y_i \leq y) = F_{Y_i}(y) = \begin{cases} \frac{1}{2}\exp(y), & y < 0 \\ 1 - \frac{1}{2}\exp(-y), & y \geq 0 \end{cases} \quad 5\text{-}3$$

The dependence model considers the asymptotics of the conditional distribution $\Pr(Y_{-i} \leq y_{-i} | Y_i = y_i$ where for $y_i \to \infty$ the increase of y_{-i} must result in non-degenerate margins. For this, assume the normalizing functions $\mathrm{a}_{|i}(y_i)$ and $b_{|i}(y_i)$ that have the same dimension as Y_{-i} and for which:

$$\lim_{y_i \to \infty} \left[\Pr\left\{ \frac{Y_{-i} - \mathrm{a}_{|i}(y_i)}{b_{|i}(y_i)} \leq z_{|i} \middle| Y_i = y_i \right\} \right] = G_{|i}(z_{|i}) \quad 5\text{-}4$$

where the limit distribution $G_{|i}$ has non-degenerate marginals $G_{j|i}$ for all $j \neq i$. The extremes dependence is the described by the semi-parametric regression model as:

$$Y_{-i} = \alpha_{|i} y_i + y_i^{\beta_{|i}} Z_{|i} \text{ for } Y_i = y_i > u_{Y_i},\ i = 1, \dots, d \quad 5\text{-}5$$

where $a_{|i}(y_i) = \alpha_{|i} y_i$ is the location function and $b_{|i}(y_i) = {y_i}^{\beta_{|i}}$ the scale function, with the vectors constants defined as $\alpha_{j|i} \in [-1,1]$ and $\beta_{j|i} \in (-\infty, 1)$ for all $j \neq i$. Detailed descriptions for the CEM can be found in Heffernan and Tawn (2004) and Keef et al. (2013).

5.3.2.3.3 ELM

The second method used to post-process the PBM simulated flow is an ELM. It is a machine learning technique developed by Huang et al. (2006) which has been applied to streamflow modelling and forecasting (e.g. Deo and Şahin, 2016; Yaseen et al., 2016). It has a simple form of one input, one hidden and one output layer and can be defined as:

$$\sum_{i=1}^{\Lambda} B_i h_i(m_i \cdot x_t + n_i) = z_t \qquad 5\text{-}6$$

where Λ is the total number of nodes, B are the estimated weights between the nodes of the hidden and output layers, and $h(m, n, x)$ is the activation function with weights $m_i \in \Re^d$, biases $n_i \in \Re$ and the explanatory variable of the training dataset $x_t \in \Re^d$. Here, i and d denote the index of a specific hidden neuron (HN) and the number of input neurons, respectively, and Z is the model output.

The input weights and hidden layer biases are chosen randomly initially and the output weights are estimated iteratively via least squares. Once the model has been trained, forecasts are obtained by introducing the testing dataset. The number of HNs in the hidden layer presents a classic problem of over-fitting and under-fitting and is commonly defined empirically (Sun et al., 2008).

5.3.2.3.4 Application

Both the CEM and ELM model were applied using a jackknife procedure (Miller, 1964). Initially, a peak flow (measured and simulated) was left out of the dataset to be used for testing, while the remainder were used for training. From the fitted CEM to the training dataset, 50,000 stochastic simulations were obtained. The realisations of the conditioning variable X_i (pseudo-PBM simulated) that were closer (<0.1) to the maximum PBM-simulated peak of the testing data were retrieved. Then, the corresponding X_j (pseudo-observations) were considered and by calculating their median value, a forecast of the maximum peak was obtained. The ELM was trained using PBM simulated data and in experimentation, measured soil moisture content as well. Using the data that were left out for testing purposes (except for the maximum), forecasts were obtained.

For each peak flow event as defined by the selected threshold (Section 5.3.2.3.1), flow values smaller than the maximum flow were forecasted by the ELM and the CEM was used only to forecast the maximum one. The CEM and ELM were both applied to the PBM simulated daily flow data while only the ELM was used to post-process the hourly-aggregated-to-daily (H2D) PBM simulations. The reason for omitting the CEM was that the H2D simulations showed an increased accuracy in simulating the maximum peaks, sometimes over-estimating them and, thus, the CEM was unnecessary. It should also be noted that SM was used only as a covariate in the ELM model. The resulting six study models are consequently referred to as Modelled Daily, Hybrid Daily, Hybrid Daily with SM, Modelled H2D, Hybrid H2D and Hybrid H2D with SM.

5.3.2.4 Models for exploratory analysis

5.3.2.4.1 2.3.1. Variograms

The temporal dependence of the measured and modelled flow was characterised by means of variograms. The variogram is a function that relates semi-variance to separation in time h (or space for spatial variables). In the context of spatial data h, which is known as the lag, is a vector describing distance and direction. For temporal data it is a scalar variable for any value of h, the empirical variogram is given by:

$$\gamma(h) = \frac{1}{2}\mathrm{E}[\{Z(t) - Z(t+h)\}^2], \qquad 5\text{-}7$$

where $Z(t)$ and $Z(t\ +\ h)$ are the values of the random function Z at time points t and $t\ +\ h$.

We estimated the values of $\gamma(h)$ by the method of moments (e.g. Webster and Oliver, 2007), which is given by:

$$\hat{\gamma}(h) = \frac{1}{2m}\sum_{i=1}^{m}[Z(t+h) - Z(t_i)]^2 \qquad 5\text{-}8$$

where $Z(t_i)$ and $Z(t_i\ +\ h)$ are the observed values at times t_i and $t_i + h$ separated by h, for $m(h)$ paired comparisons at that lag. As observations of the process become further apart (quantified by h) they typically become less correlated, and often there exists a lag beyond which there is no correlation.

We fitted plausible models to the empirical variograms using the directive FITNONLINEAR in GenStat (v. 18) (Payne et al., 2008). Authorised variogram models have simple shapes, but can be combined additively to represent more complex shapes (Webster and Oliver, 2007).

The base variogram models that we considered were spherical, circular and exponential (see S1 for details).

In this research, we computed empirical and modelled variograms for measured flow, measured precipitation and measured SM together with the simulated flow data from each of the six models described above. For measured and modelled flow and precipitation, data were log transformed before variograms were fitted because of the skew in the data (i.e., transforms were used to facilitate authorised variogram model fits). The use of transformed data will have a clear bearing on the interpretation of the variograms compared to variograms constructed from untransformed data. This data pre-processing decision (for the variography only) is reviewed in the discussion.

5.3.2.4.2 Wavelet analysis

We used the maximum overlap discrete wavelet transform (MODWT) to analyse (Percival and Walden, 2000) the performance of each model. The wavelet transform comprises a set of basis functions which can be convolved with a series of data to produce wavelet coefficients. Each basis function has, what is known as compact support, which means that it is non-zero for only a finite period. This property means that convolution with a wavelet basis function picks up localised features in the data, unlike a Fourier transform which extracts information on a frequency component across the whole series. The set of basis functions are all dilations and translations of a basic wavelet function known as the mother wavelet. For the MODWT the function is translated by unit steps across the series, and dilated by a scale parameter, a_j, which increases in a dyadic (power of two) sequence $a_j = 2^j t$ $(j = 1,2, \dots J)$ and where t is

the sample interval of the time-series. The maximum dilation J must satisfy $n \geq 2^J$, where *n* is the length of the time-series.

The wavelet coefficients calculated using a basis function with dilation a_j are nominally associated with the scale interval $[2^j, 2^{j+1}]$ (Percival and Walden, 2000), and their locations relate to the location of the non-zero part of the basis function. A scaling function associated with the mother wavelet function completes the set of basis functions. When the time-series is convolved with the scaling function a set of approximation coefficients (or scaling coefficients) are produced. These are related to the mean of the time-series.

The wavelet transform is invertible, that is to say, that a complete set of wavelet and approximation coefficients can be used to reconstruct the original signal. If all the coefficients are set to zero except those from a particular scale and these are then back transformed the result is the component of the original time-series that is associated with that scale. In this way, a set of components, one for each of the scale intervals defined and one associated with the approximation coefficients, can be obtained. This is known as a multi-resolution analysis (MRA). The original time-series is given by the sum of the components.

As well as decomposing the signal into scale components, the wavelet coefficients can be used to calculate scale-specific components of the variance in the signal, known as wavelet variances. The wavelet variance for the scale $2^j x$ is computed by

$$\sigma_{u,j}^2 = \frac{1}{2^j n_j} \sum_{k=1}^{n_j} \{d_{j,k}^u\}^2, \qquad 5\text{-}9$$

where $d_{j,k}^{u}$is the *k*th MODWT coefficient of time-series variable *u* at scale $2^{j}x$ (Percival and Walden, 2000), and n_j is the number of wavelet coefficients calculated at the jth scale (for details see Milne et al., 2009).

Similarly, given two signals, u and v, a wavelet correlation for each scale interval can be computed. This is given by

$$\rho_{u,v,j} = \frac{C_{u,v,j}}{\sigma_{u,j}\sigma_{v,j}}. \qquad 5\text{-}10$$

where $C_{u,v,j}$ is the wavelet covariance between the two variables and is given by

$$C_{u,v,j} = \frac{1}{2^{j}n_j}\sum_{k=1}^{n_j} d_{j,k}^{u} d_{j,k}^{v}. \qquad 5\text{-}11$$

These formulae give the wavelet correlation and wavelet variance over the entire time-series. Unlike the variogram, however, a key feature of the wavelet transform, is that it captures local variation. It is possible therefore possible to test for significant changes in the wavelet variance and correlation at each scale (Lark and Webster, 2001).

In this research, we used Daubechies's extremal phase wavelet (Daubechies, 1988) with two vanishing moments, since this has a very compact support, and a maximum dilation of eight to investigate model performance across scales. We first computed the wavelet variance for modelled flow data using time-series from each of the six models described above. We then compared the partition of variation across the scales to see which of our models captured the behaviour observed in the measured data. Similarly, we computed the wavelet correlation between modelled and measured data to determine which scales performed best.

5.3.2.4.3 Change point detection with wavelets

To determine how the models performed over time and to see if there were significant changes in performance, we conducted an MRA of the residuals between the modelled and measured flows and determined the significant change points.

Finally, and of key interest here, is the concept of identifying extreme events from model predictions. Therefore, we also explored variance change point detection for the Modelled Daily and the Modelled H2D outputs to evaluate whether the onset of extreme events observed in the measured flow data was reflected in the model-based analysis. Note that we did not do this for the hybrid models because part of their construction is based on defining when extreme events occur.

5.4 Results

5.4.1 Time-series and model predictive performance

All six models captured well the general pattern and the peaks of the measured flow (Figure 5-3 and Figure 5-4). Scatterplots of measured flow against simulated flow, and the associated linear correlations, are presented in Figure 5-5 to provide a detailed evaluation on the performance of each model. The Modelled H2D and the two Hybrid H2D models produced the largest correlation coefficient with measured flow, followed by the Modelled Daily and the Hybrid Daily models. Adding SM as a covariate did not increase model accuracy. The scatterplots also indicate that all the H2D-based models are more accurate in terms of high flows as they are closer to the 1-1 line compared to all the Daily-based models. This is confirmed with larger correlations. Surprisingly, the smallest correlations exist between the Hybrid Daily with SM and all the H2D-based models.

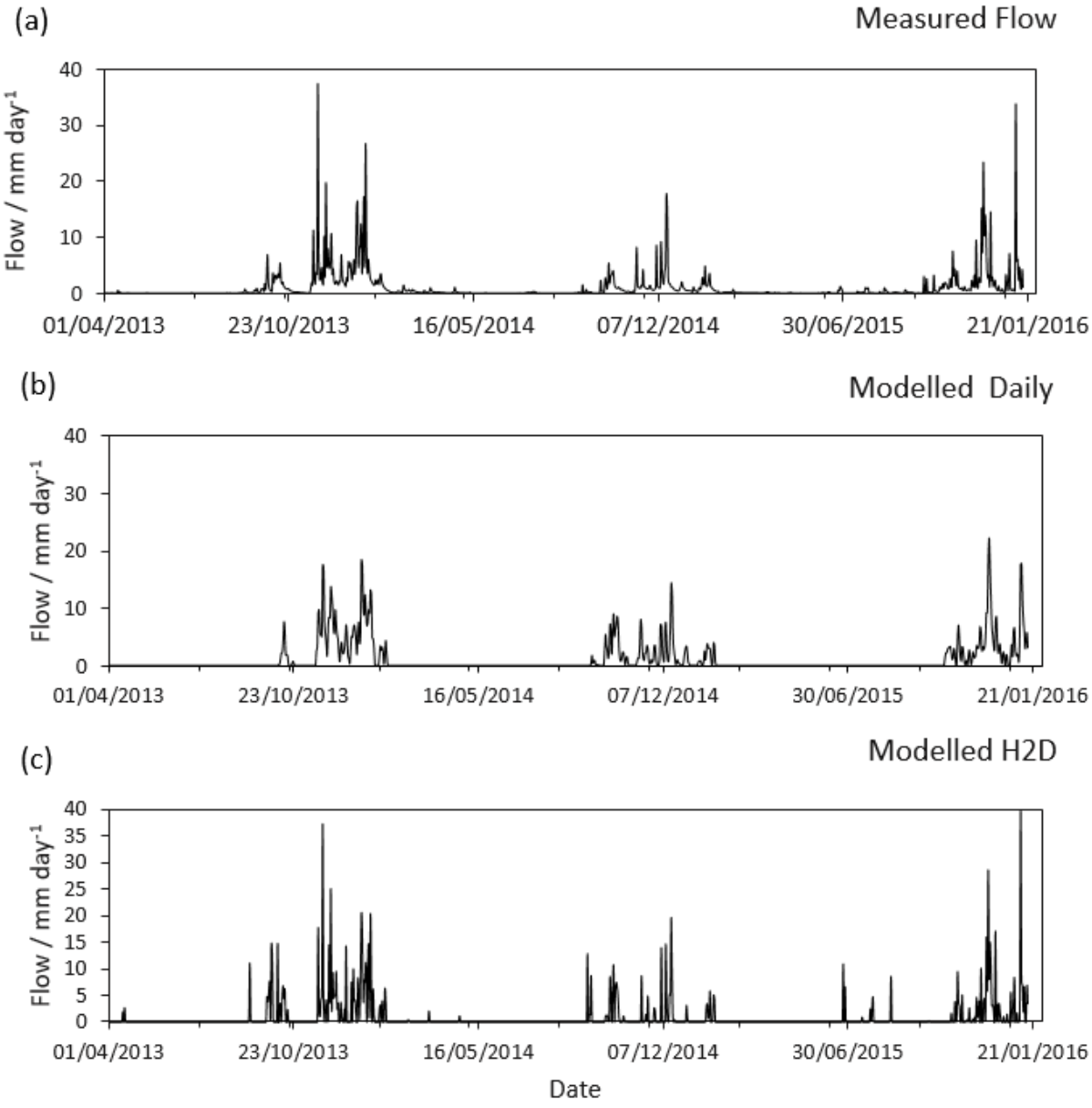


Figure 5-3: (a) Daily measured flow,(b) PBM simulated flow at daily resolution (Modelled Daily) and (c) simulated at hourly resolution aggregated to daily (Modelled H2D).

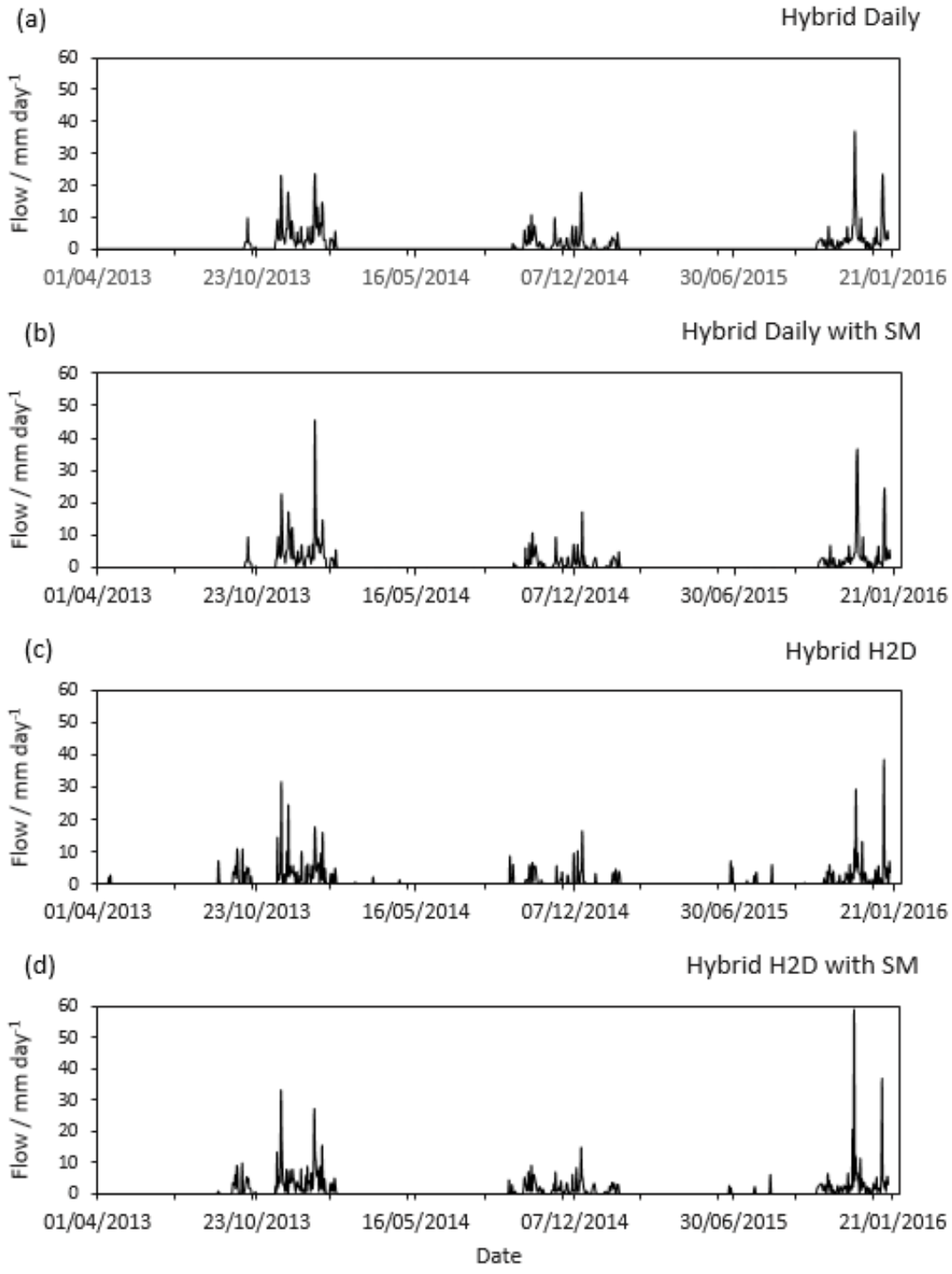


Figure 5-4: Hybrid models a) with CEM applied to the maximum daily PBM simulated flow within a peak event and ELM to all other points in the peak event, b) as in (a) but with soil moisture (SM) as a covariate in the ELM model, c) with ELM only applied to the hourly PBM

simulated and aggregated to daily flow, d) as in (c) but with SM as a covariate.

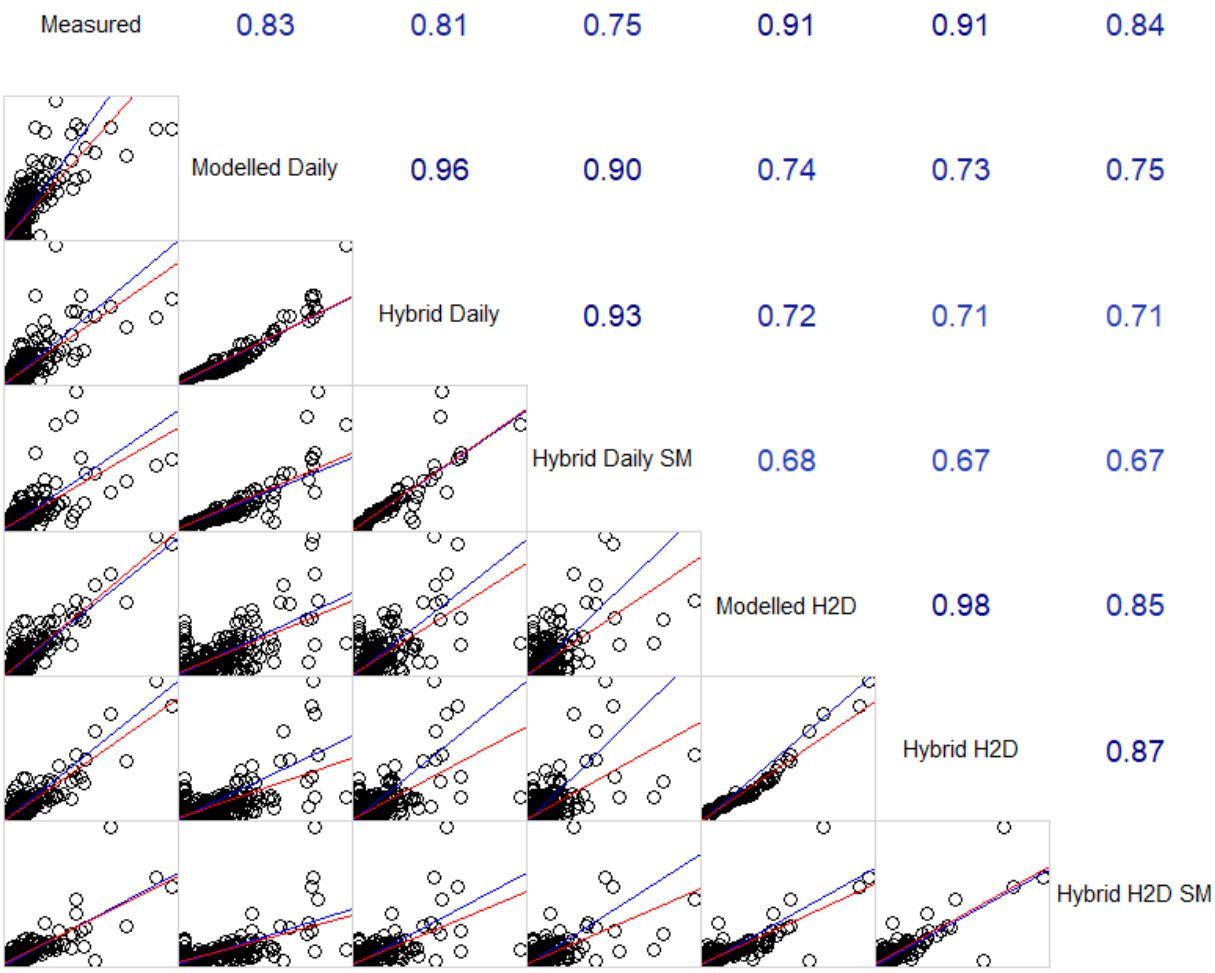


Figure 5-5: (Bottom left) scatterplots with 1:1 (blue) and regression (red) lines and (top right) correlations between measured and simulated flow and between flow simulations from the models only.

5.4.2 Variograms

Empirical variograms were computed for the three measured variables (flow, precipitation and SM) and six simulated flow variables. Only the SM variable remained in un-transformed space, while the rest were log-transformed to facilitate the identification of clear structures in the respective autocorrelated processes. Authorised variogram models could be fitted to all empirical variograms except for measured flow and SM (Figure 5-6). This was due to a concave upwards behaviour at certain lag ranges in the respective empirical variograms. In all

cases, a double spherical model fitted best indicating a clear nested structure with two scales of temporal variation. A nested characteristic was also broadly apparent in the un-modelled empirical variograms of measured flow and SM.

Across the variograms for measured and simulated flow there was a consistent short-range component with range parameter of approximately 12 days and a longer-range component of around 180 days (see Appendix D for the variogram model parameters). The short-range component accords with that seen in the precipitation data. However, for SM, the short- and long-range components were approximately 20 and 175 days. The relatively large nugget in the precipitation variogram suggests that there is little temporal autocorrelation in the data. The gradual increase in variance for lags in the range of 12-180 days indicates that the autocorrelation in the data decreases approximately linearly.

Variograms for modelled flow were compared with the empirical variogram for measured flow (Figure 5-6d-i). Good correspondence with measured flow was found for the Modelled Daily, Hybrid Daily, Hybrid Daily with SM and Hybrid H2D with SM model outputs. Thus, simulations from only four of the six models broadly captured the observed autocorrelation in the measured flow data. It is notable that the otherwise poorly fitting H2D-based model was improved in this respect by the use of SM as a covariate.

Variograms depicted in Figure 5-6 d, e and f, have similar characteristics and this is driven by the fact that they are based on the same underlying model output (Modelled Daily). The Hybrid Daily data show a small decrease in the overall sill and the addition of SM makes negligible difference to the variogram. The application of the hybrid model with the H2D has a more significant impact than on the Daily Modelled data, which is confirmed by the change

in the parameter estimates (S1). Furthermore, SM significantly changes the variance of the H2D Modelled data, which becomes similar to all the Daily Modelled.

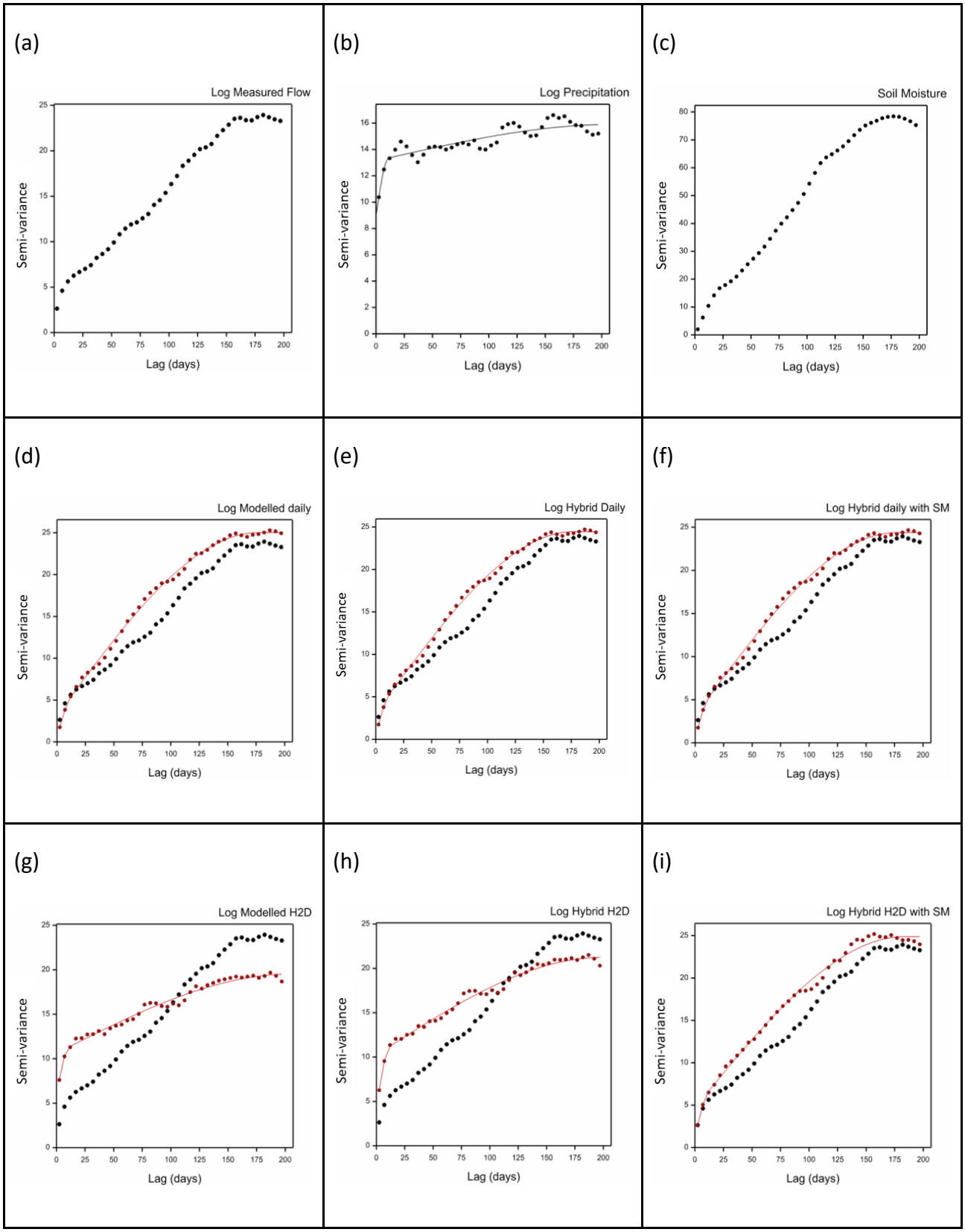


Figure 5-6: Empirical variograms of measured (a) log flow, (b) log precipitation and (c) soil moisture. The black line shows the variogram model fitted to the measured data (for precipitation only). Subplots (d-i) show the empirical variograms for log modelled flow variables (red disks) with their respective fitted variogram models (red line) and the empirical variograms of measured log flow for comparison (black disks).

5.4.3 Wavelet Analysis

5.4.3.1 Wavelet variance

The wavelet variance results are given in Figure 5-7 and Figure 5-8. The partition of wavelet variance in the measured discharge data shows that the largest component exists at the finest scale (2-4 days). The variance then falls sharply, with a small peak at the 32-to-64 day scale. It then increases with scale, with the coarsest scale relating to annual variation (Figure 5-7a). Comparing the wavelet variance of the measured data with the PBM simulations (Figure 5-7b) shows that Modelled H2D overestimates the fine-scale wavelet variance and Modelled Daily underestimates it. At coarser scales, the variance of Modelled H2D becomes similar to the measured one while the Modelled Daily deviates, suggesting that coarse scale variation is overestimated. Similar to the measured flow and Modelled H2D, precipitation shows the greatest variation at the fine scale. Conversely, the wavelet variance for SM increases broadly with scale, which reflects the fact that the processes controlling it dampen the fine-scale variation relative to the coarse scale. The Hybrid H2D model captures best the measured wavelet variance at scales finer than 32 days (Figure 5-8). At coarser scales the Hybrid H2D with SM performs best in this respect (Figure 5-8). Using SM as a covariate in the Hybrid Daily model does not increase the accuracy of the predicted variation at coarser scales, however (Figure 5-8).

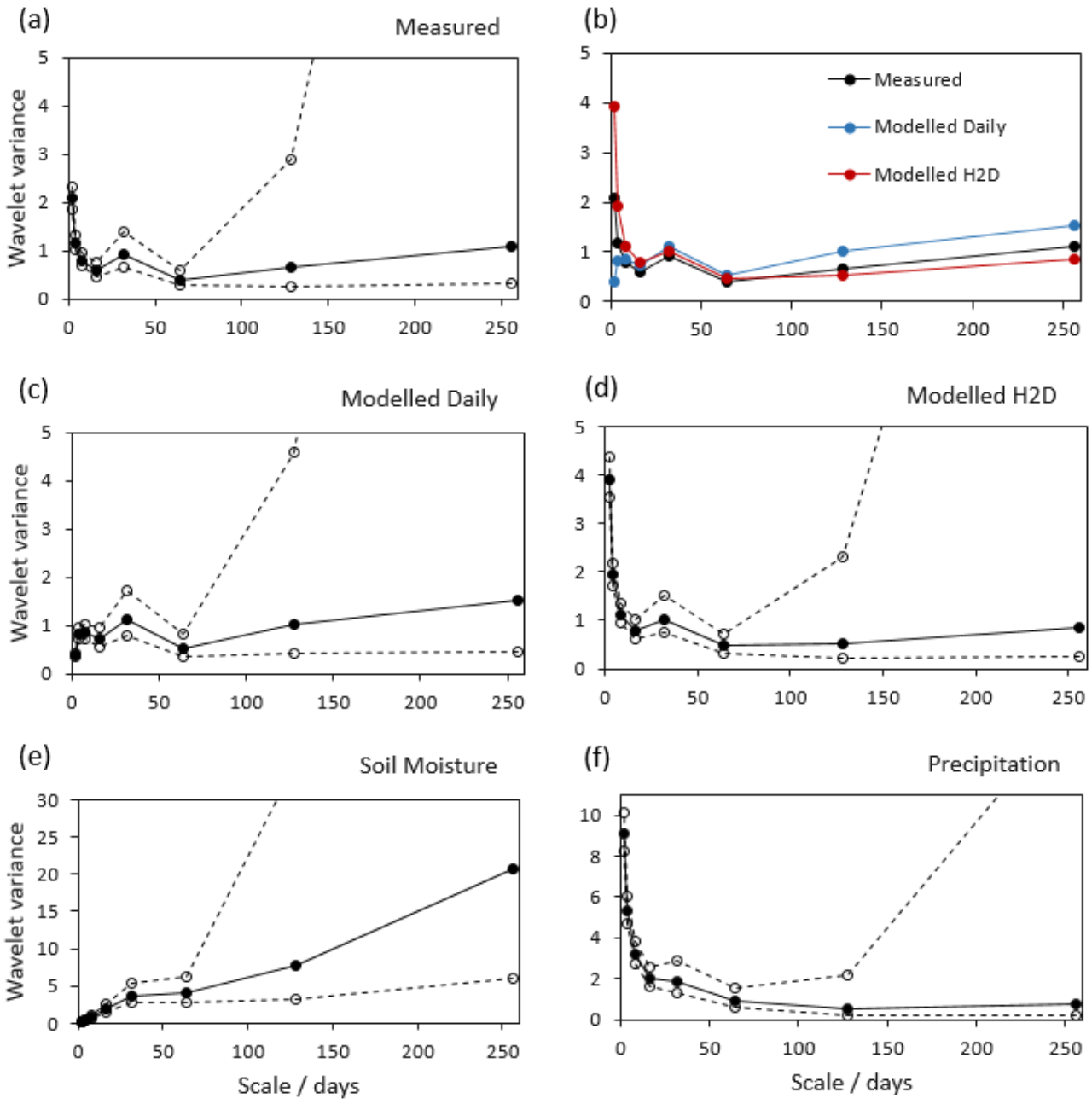


Figure 5-7: The wavelet variance for measured (plots a, e, f for flow, SM and precipitation, respectively) and modelled (c and d for Daily and N2D, respectively) data. The wavelet variance is given by the solid discs which mark the lower bound of the scale interval that each wavelet variance is associated with. The open discs show the 95% confidence intervals. The lines are given to aid the eye. Plot (b) compares measured with modelled flow on the same plot.

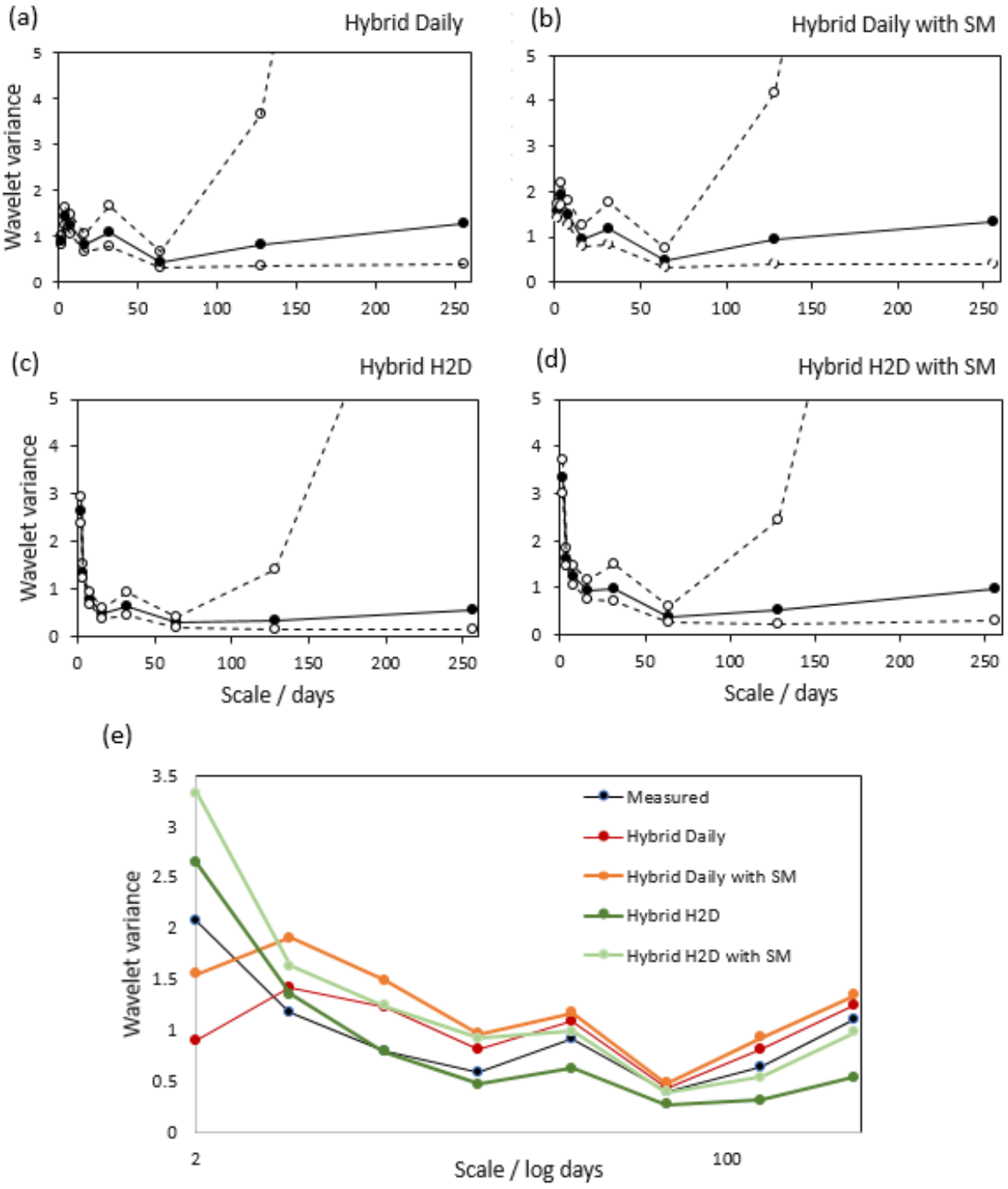


Figure 5-8: The wavelet variance for flow simulated with each of the hybrid models. The wavelet variance is given by the solid discs which mark the lower bound of the scale interval that each wavelet variance is associated with. The open discs show the 95% confidence intervals. The lines are given to aid the eye. The bottom plot shows the wavelet variance for all of the hybrid models plotted together with the wavelet variance for the measured data. The scale is presented on the log scale (base 10) to aid inspection of the finer scale variances.

5.4.3.2 MRA of residuals

The MRAs of the residuals for each of the six models are shown in Figure 5-9. The significant changes in model performance (as indicated by the red vertical lines) show that the models all start to fail around the three large bursts of flow activity (we note that for clarity we omitted change points on the two fine-scale components where changes were numerous). All of the models capture the coarse scale variation well (as demonstrated by the near flat variation in the top three variance components). Over the whole time period, residual variation is smallest for the Hybrid H2D at the finer scales and for Hybrid models with SM at the coarsest scales (Table 5-1).

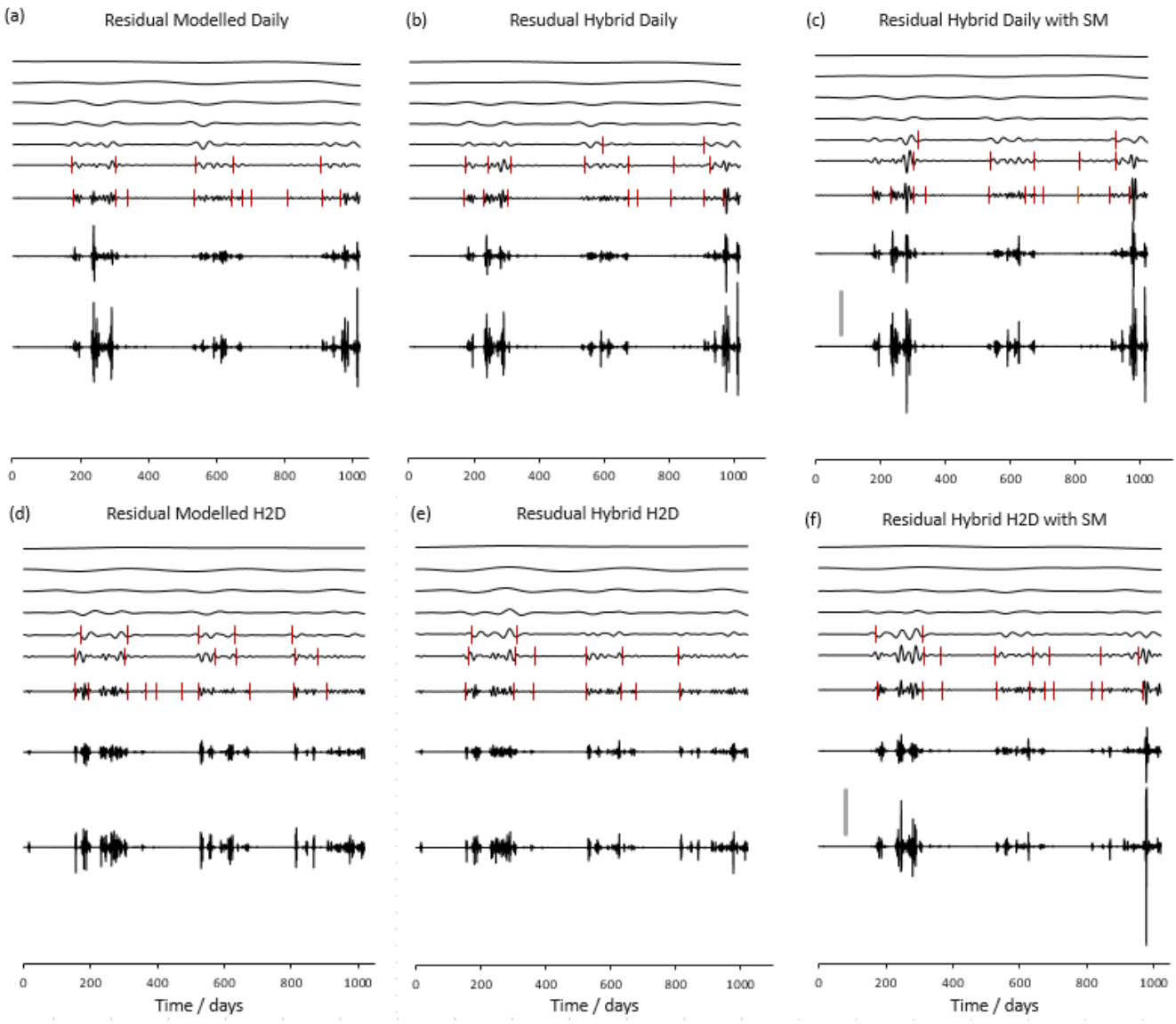


Figure 5-9: The MRA for the residuals of each model considered shown as stacked plots. The approximation component is shown at the top of each subplot with variance components plotted below from coarsest at the top to finest at the bottom. The solid grey bar indicates a 10-unit scale which is common across all subplots. The wavelet variances of each component are given in Table 5-1. We note that because the top component is the approximation component it does not have an associated wavelet variance. Significant change points in the residual variance are shown by the red vertical lines. These are only shown for scales above 8 days.

Table 5-1: The wavelet variances of the residuals for each model.

	Modelled daily	Hybrid daily	Hybrid daily with SM	Modelled H2D	Hybrid H2D	Hybrid H2D with SM
Scale 256 - 512	0.064	0.035	0.033	0.058	0.134	0.049
Scale 128 - 256	0.094	0.050	0.069	0.067	0.112	0.048
Scale 64 - 128	0.071	0.047	0.061	0.057	0.061	0.049
Scale 32 - 64	0.095	0.079	0.138	0.104	0.112	0.157
Scale 16 - 32	0.112	0.153	0.325	0.173	0.142	0.300
Scale 8 - 16	0.210	0.323	0.623	0.211	0.156	0.328
Scale 4 - 8	0.500	0.685	1.109	0.428	0.266	0.657
Scale 2 - 4	1.533	1.778	2.390	0.915	0.489	1.541

5.4.3.3 Wavelet correlations

Across the scales, the models derived from PBM flow simulations at the hourly resolution (Modelled H2D, Hybrid H2D and Hybrid H2D with SM) produce a large wavelet correlation with measured flow (>0.7) whereas those based on simulated flow at the daily resolution are less correlated at finer scales (Figure 5-10). Hybrid H2D is the best performing model at finer scales (<32 days), while at coarser scales (32> days) all models produce a large correlation with the measured data. Surprisingly, the Hybrid Daily model has a smaller fine-scale

correlation with the measured data than the Modelled Daily model. Using the SM covariate increases the coarse scale correlations only marginally.

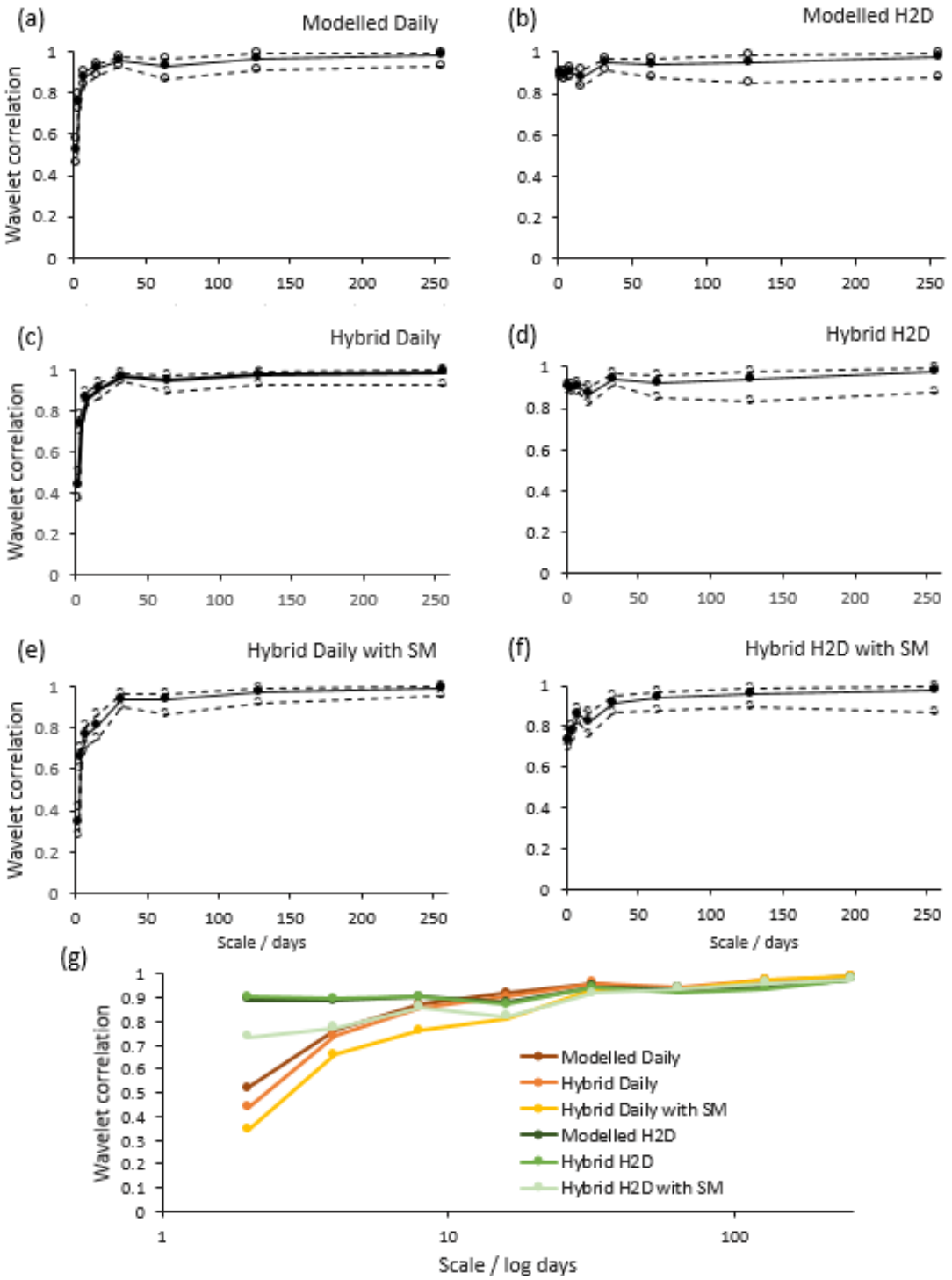


Figure 5-10: The wavelet correlation between simulated and measured flow data. The wavelet correlation is given by the solid discs which mark the lower bound of the scale interval to which each wavelet correlation is associated. The open discs show the 95% confidence intervals. The lines are

given to aid the eye. The bottom plot shows the wavelet correlation for all models plotted together. The scale is presented on the log scale (base 10) to aid inspection of the finer scale correlations.

5.4.3.4 Wavelet Analysis for detection of extreme events

The MRA and wavelet variance change point detection shows that broadly, the two PBM simulation models (Modelled Daily and Modelled H2D) capture the significant changes in variance at each scale. This is demonstrated by the similarity in the location of change detection points between modelled and measured flow (Figure 5-11a–c). There is a small burst of activity at just after 800 days which is detected in the 8-to-16 day scale component of the measured data that is not captured in Modelled Daily but is overestimated by the Modelled H2D. The magnitude of the estimated local wavelet variance is related to the likely number of extreme (peak flow) events and how soon an extreme event is likely to occur (Figure 5-11d–e) (see Appendix D for Modelled H2D).

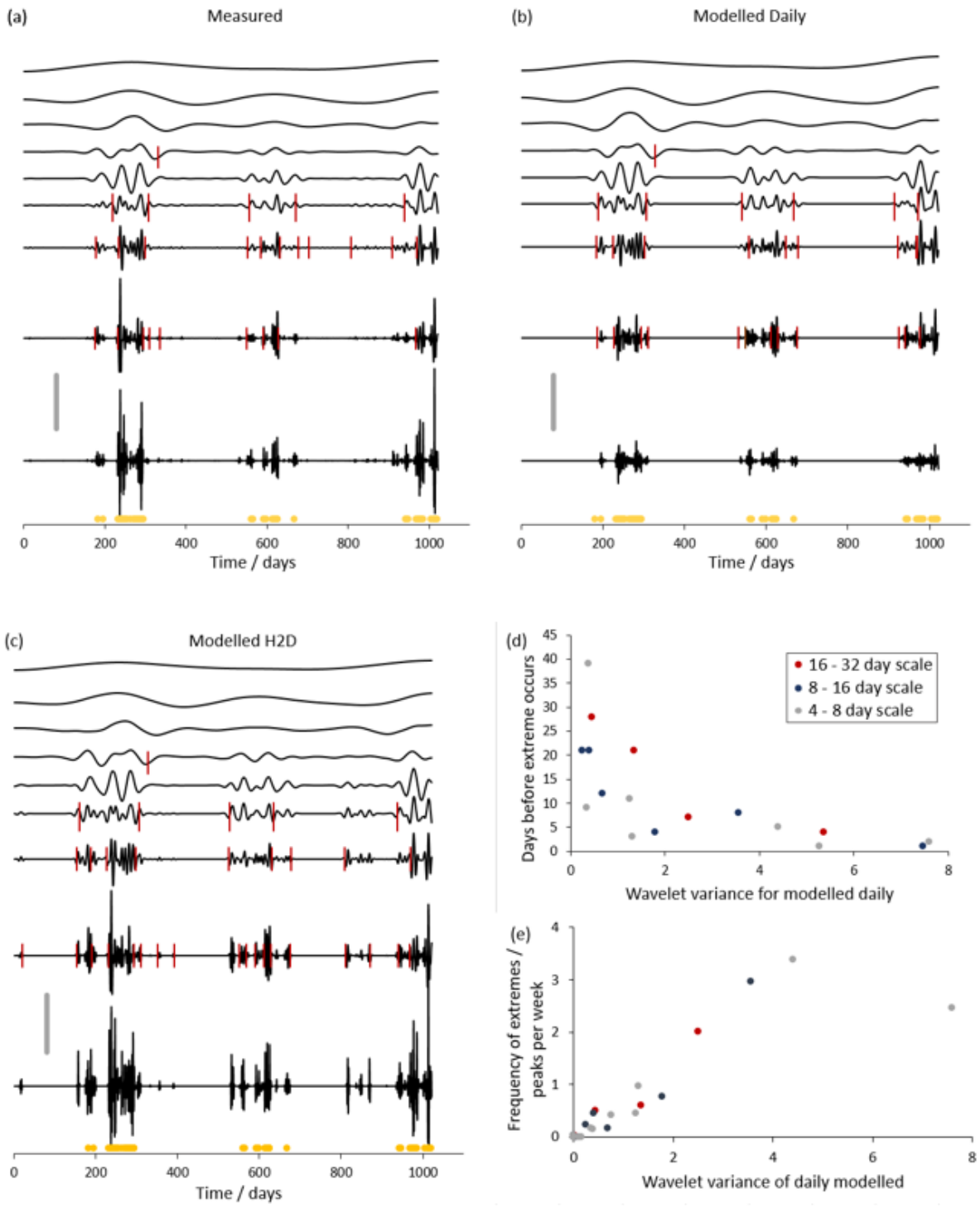


Figure 5-11: The MRAs of (a) measured flow, (b) Modelled Daily flow and (c) Modelled H2D flow shown in stacked plots. The approximation component is shown at the top of each subplot with variance components plotted below from coarsest at the top to finest at the bottom. The solid grey scale bar indicates 10 units. Significant change points in the residual variance are shown by the red vertical lines. These are not shown for scales above 4 days. The yellow dots indicate the extremes (peak flows) as detected by the peaks over threshold method for the measured data (Curceac et al.

2020a; 2020b). Plot (d) shows the relationship between the number of days after a change point that an extreme value is detected and the local wavelet variance and (e) the frequency of extremes and the local wavelet variance.

5.5 Discussion

Accurate modelling and forecasting of runoff from agricultural land is important for management of nutrient losses and water pollution. In the context of grassland agriculture, water flow is most commonly modelled using process-based models. However, recent advances suggest that a hybrid modelling approach combining statistical distributions and machine learning can increase predictive power. In this research, we presented and evaluated six alternative models for predicting flow data, all variations on the same PBM (SPACSYS); three were existing models (Modelled Daily, Modelled H2D, Hybrid Daily), while three were new models (Hybrid Daily with SM, Hybrid H2D, Hybrid H2D with SM).

A simple correlation analysis (Figure 5-5) indicated that Modelled H2D and Hybrid H2D were the most accurate predictors of water flow (both yielding correlations of r = 0.91), where surprisingly, the inclusion of SM provided no additional predictive information. The fact that SM has no positive effect on the models' performance could have several explanations. Measuring SM is known to be more difficult compared, for example, to measuring precipitation. Therefore, a greater uncertainty in the SM measurements is likely (see below). Moreover, the flow is representative over the whole sub-catchment gathered at the flume whereas SM (and precipitation) is measured at only one point and at only one depth of 10 cm and thus, may not be as representative of catchment-scale SM. The relatively poor model performance may also result from overfitting to the training dataset. Another possible explanation could be the fact that SM is already taken into account as a predicted internal state variable in the PBM, which captures the seasonality.

The H2D-based models were considered more accurate in terms of predicting the extremes. All six models could characterise the coarse-scale (or global) behaviour in flow and had reasonable predictive power. Given that our focus is on the prediction of extreme events and identifying the scales associated with these events, it was necessary to evaluate model performance across scales. Therefore, we explored using diagnostics that are able to reveal how well a model is able to capture the scale dependence in the observed behaviour (wavelet analysis) and how well structural auto-correlation is preserved (variography). Variograms provide a broad, global assessment, while wavelets provide a detailed, local assessment. This combined approach may be regarded as complementary to assessments undertaken more routinely based on model prediction accuracy provided through various accuracy metrics (Smith et al., 1997), which can similarly be transferred to a detailed, local form (Harris et al., 2013; Comber et al., 2017; Tsutsumida et al., 2019).

Note that for variography we chose to log transform the measured and modelled flow data. Asymmetry or skewness in data generally has little effect on variogram estimation for large samples, and so predictions can usually be done safely with the raw data (Webster and Oliver, 2007). However, in our case we found that a “hole effect” in the empirical variogram meant it was not possible to fit a valid variogram model (known as an authorised model in geostatistical literature) without transformation. Transformation does, however, dampen the extremes in the data. Therefore, for variography we compared only the variogram models between measured and modelled variants. First, comparing the variograms of the modelled and measured flow data (Figure 5-6) it is evident that the temporal autocorrelation at shorter lag times (approximately less than 70 days) is not captured well by Modelled H2D and Hybrid H2D (our best predictive models from above). This relates to a tendency to over-predict fine-

scale variation in flow. In each case, the double spherical model was found to be the best fitting model suggesting that there exist two substantial sources of variation in the data. All modelled variograms could capture the short- (approx. 12-days) and the long-range processes (approx. 185 days) observed in the measured flow data. The former accords with the short-range process observed in the measured precipitation data, a time-scale at which the Madden-Julian Oscillation (MJO) influences the North Atlantic weather regimes (10-12 days, Met Office, UK). The long-term process, which is approximately half a year, is likely to relate to seasonal variation.

It is clear from the scatterplots of Figure 5-5 that there are issues of under-prediction of peak flows associated with models derived from the Daily PBM simulation. This is reflected in the wavelet variance where it is evident that the fine-scale wavelet variance is underestimated (Figure 5-7b). The hybrid approach mitigates this effect to some extent, but variation is still smaller than it should be at the fine scale (bottom plot in Figure 5-8). In all three Daily-based models, the relatively small fine-scale wavelet variation is overcompensated for at mid-to-coarse scales. Conversely, the H2D-based models tend to overestimate fine-scale variation (Figure 5-7b and Figure 5-8) with the most extreme effects seen in Modelled H2D (Figure 5-7 b). The hybrid models dampen this overestimation in the H2D-based models with Hybrid H2D capturing the fine-scale variation the best out of all six models. Hybrid H2D also shows the overall best wavelet correlation at finer scales (<32 days), while at coarser scales (32> days) all models produce a large wavelet correlation with the measured data (Figure 5-10). Thus, for Hybrid H2D, this complements the high performance of its standard correlation with the measured data.

A key advantage of wavelets is their ability to capture local behaviour. In terms of model behaviour, we used the approach proposed by Rust et al. (2014) and inspected the model residuals using a MRA (Figure 5-9). It is evident from the residuals that the model performance is not consistent across time and that, in particular, the Daily-based models and Hybrid H2D with SM perform particularly poorly over the last major burst of activity (900 days onward). This corresponds with a period where the soil is quite wet (close to saturated) according to the measured data and so suggests that this local measurement of soil moisture and the daily modelled predictions do not capture the more complex soil-water dynamics that operate across the sub-catchment. Interestingly, the H2D models without the SM covariate do not produce a similar issue. Except for the temporal variability of SM, the spatial variability (which is not accounted for as SM is measured at one location only) could affect the performance of the models. However, the variability in SM content would not be expected to vary significantly across a field such as the one used in this study. The soil properties are relatively constant across the field, there is no interaction with other surface or underground water systems, the vegetation is homogenous as it consists of grass only and it is safe to assume that the precipitation is evenly distributed across such a small filed. The only possible source of SM spatial variation would be due to topography, where parts of the field at a lower elevation would be expected to have increased water content compared to locations at higher altitudes due to gravitational forces.

The extreme events identified using the automated threshold stability method (as given in Curceac et al., 2020a; 2020b) did accord somewhat with the wavelet change point detection analysis (Figure 5-11). The local wavelet variance of the model predictions (i.e., only those from the Modelled Daily and Modelled H2D) was correlated with the number of extreme

events and a large wavelet variance suggested that extreme events were imminent. The wavelet-based method was less efficient for predicting extremes, than simply applying the automated threshold stability method to the model prediction. However, it serves well in an exploratory and complementary context.

5.6 Conclusions

In this research, we demonstrated that the dual use of a variogram and wavelet analysis could provide a useful exploratory assessment of existing and newly proposed hydrological models, with respect to how they capture changes in flow variance at different scales and how this correlates with measured flow; all in the context of capturing extreme flow events. Variograms provided a broad, global assessment, while wavelets provided a detailed, local assessment, both of which would complement standard assessments based only on prediction accuracy. In doing so, a more complete understanding of model behaviour and model performance was elucidated.

Such detailed assessments are particularly important for hybrid models which not only depend on the parameterisation of the underlying process-based model component (and its data requirements), but also the accurate estimation of the parameters of the statistical data-driven component(s) (in this case for the characterisation of extreme flows). Although study models benefitted from fine-resolution measured data from an agricultural research platform, such data are increasingly becoming routine in water monitoring, entailing our complex hybrids and our involved methods of assessment should increasingly become the norm given a hybrid model should increase the accuracy of simulating peak flows over a process-based model alone. This is to be welcomed given the drivers of climate change and

changing patterns of rainfall are complex and so evaluating the risk of extreme water flows and associated flooding will continue to require complex solutions.

Authors contribution statement

Stelian Curceac 60%: Conceptualisation, Methodology, Software, Formal analysis, Writing-Original Draft, Writing-Review & Editing.

Alice Milne 25%: Conceptualisation, Methodology, Software, Formal analysis, Writing-Original Draft, Writing-Review & Editing, Supervision.

Peter Atkinson 5%: Conceptualisation, Writing-Review & Editing, Supervision, Funding acquisition.

Lianhai Wu 5%: Conceptualisation, Writing-Review & Editing, Supervision.

Paul Harris 5%: Conceptualisation, Data curation, Writing-Review & Editing, Supervision, Funding acquisition.

Acknowledgements

Rothamsted Research receives grant aided support from the Biotechnology and Biological Sciences Research Council (BBSRC) of the United Kingdom. This research was funded by Rothamsted Research and Lancaster Environment Centre, the BBSRC Institute Strategic Programme (ISP) grant, “Soils to Nutrition” (S2N) grant numbers BBS/E/C/000I0320, BBS/E/C/000I0330 and the BBSRC National Capability grant for the North Wyke Farm Platform grant number BBS/E/C/000J0100.

The study datasets are freely available from https://www.rothamsted.ac.uk/north-wyke-farm-platform and the SPACSYS (PBM) model can be found here https://www.rothamsted.ac.uk/rothamsted-spacsys-model. The variogram analysis was

conducted in GENSTAT (VSN International, 2019), while R software (R Core Team, 2019) was used for the implementation of the hybrid models, where the CEM is from the texmex R package (Southworth et al., 2020) and the ELM from the elmNNRcpp R package (Mouselimis and Gosso, 2020).

References

Bates, B. C., Kundzewicz, Z. W., Wu, S. and Palutikof, J. P. (2008). Climate Change and Water. Technical Paper of the Intergovernmental Panel on Climate Change, IPCC Secretariat, Geneva, 210 pp.

Bouraoui, F., Grizzetti, B., Granlund, K., Rekolainen, S. and Bidoglio, G. (2004). Impact of Climate Change on the Water Cycle and Nutrient Losses in a Finnish Catchment, Climatic Change, 66(1-2), 109-126. doi: 10.1023/B:CLIM.0000043147.09365.e3.

Brown, I., Bardgett, R., Berry, P., Crute, I., Morison, J., Morecroft, M., Pinnegar, J., Reeder, T. and Topp. K. (2016). UK Climate Change Risk Assessment, Chapter 3: Natural Environment and Natural Assets.

Chilès, J. P. and Delfiner P. (2009). Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons.

Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer, London, UK.

Comber, A., Brunsdon, C., Charlton, M. and Harris, P. (2017). Geographically Weighted Correspondence Matrices for Local Error Reporting and Change Analyses: Mapping the Spatial

Distribution of Errors and Change, Remote Sensing Letters, 8(3), 234-243. doi: 10.1080/2150704X.2016.1258126.

Curceac, S., Atkinson, P. M., Milne, A, Wu, L. and Harris, P. (2020a). Adjusting for Conditional Bias in Process Model Simulations of Hydrological Extremes: An Experiment Using the North Wyke Farm Platform, Frontiers in Artificial Intelligence, 3. doi: 10.3389/frai.2020.565859.

Curceac, S., Atkinson, P. M., Milne, A., Wu, L. and Harris, P. (2020b). An Evaluation of Automated GPD Threshold Selection Methods for Hydrological Extremes across Different Scales, Journal of Hydrology, 585, 124845. doi: 10.1016/j.jhydrol.2020.124845.

Daubechies, I. (1988). Orthonormal Bases of Compactly Supported Wavelets, Communications on Pure and Applied Mathematics, 41(7), 909-996. doi: 10.1002/cpa.3160410705.

Deo, R. C. and Şahin, M. (2016). An Extreme Learning Machine Model for the Simulation of Monthly Mean Streamflow Water Level in Eastern Queensland, Environmental Monitoring and Assessment, 188, 90. doi: 10.1007/s10661-016-5094-9.

Field, C. B., Barros, V., Stocker, T. F. and Dahe, Q. (2012). Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change, Cambridge, Cambridge University Press. doi: 10.1017/CBO9781139177245.

Goovaerts, P. 1997. Geostatistics for Natural Resource Evaluation, Technometrics, Vol. 42.

Gringarten, E. and Deutsch, C. V. (2001). Teacher's Aide Variogram Interpretation and Modeling, Mathematical Geology, 33(4), 507-534. doi: 10.1023/A:1011093014141.

Harris, P., Brunsdon, C. and Charlton, M. (2013). The Comap as a Diagnostic Tool for Non-Stationary Kriging Models, International Journal of Geographical Information Science, 27(3), 511-541. doi: 10.1080/13658816.2012.698014.

Heffernan, J. E. and Tawn, J. A. (2004). A Conditional Approach for Multivariate Extreme Values (with Discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(3), 497-546. doi.org: 10.1111/j.1467-9868.2004.02050.x.

Huang, G. B., Zhu, Q. Y. and Siew, C. K. (2006). Extreme Learning Machine: Theory and Applications, Neurocomputing, Neural Networks, 70(1), 489-501. doi: 10.1016/j.neucom.2005.12.126.

Jaiswal, R. K., Ali, S. and Bharti, B. (2020). Comparative Evaluation of Conceptual and Physical Rainfall–Runoff Models, Applied Water Science, 10(1), 48. doi: 10.1007/s13201-019-1122-6.

Keef, C., Papastathopoulos, I. and Tawn, J. A. (2013). Estimation of the Conditional Distribution of a Multivariate Variable given That One of Its Components Is Large: Additional Constraints for the Heffernan and Tawn Model, Journal of Multivariate Analysis, 115, 396-404. doi: 10.1016/j.jmva.2012.10.012.

Kundzewicz, Z. W., Mata, L. J., Arnell, N. W., Doll, P., Kabat, P., Jimenez, B. et al. (2007). Freshwater Resources and Their Management. In Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by M. L. Parry, O. F. Canziani, J. P. Palutikof, P. J. van der Linden, and C. E. Hanson, 173–210. Cambridge University Press.

Lark, R. M., and Webster, R. (1999). Analysis and Elucidation of Soil Variation Using Wavelets, European Journal of Soil Science, 50(2), 185-206. doi: 10.1046/j.1365-2389.1999.t01-1-00234.x.

Lark, R. M., and Webster, R. (2001). Changes in Variance and Correlation of Soil Properties with Scale and Location: Analysis Using an Adapted Maximal Overlap Discrete Wavelet Transform, European Journal of Soil Science, 52(4), 547-562. doi: 10.1046/j.1365-2389.2001.00420.x.

Liu, Y., Li, Y., Harris, P., Cardenas, L. M., Dunn, R. M., Sint, H., Murray, P. J., Lee, M. R. F. and Wu, L. (2018). Modelling Field Scale Spatial Variation in Water Run-off, Soil Moisture, N2O Emissions and Herbage Biomass of a Grazed Pasture Using the SPACSYS Model, Geoderma, 315, 49-58. doi: 10.1016/j.geoderma.2017.11.029.

Miller, R. G. (1964). A Trustworthy Jackknife, The Annals of Mathematical Statistics, 35(4), 1594-1605. doi: 10.1214/aoms/1177700384.

Milne, A. E., Macleod, C. J. A., Haygarth, P. M., Hawkins, J. M. B. and Lark, R. M. (2009). The Wavelet Packet Transform: A Technique for Investigating Temporal Variation of River Water Solutes, Journal of Hydrology, 379(1), 1-19. doi: 10.1016/j.jhydrol.2009.09.038.

Mouselimis, L. and Gosso, A. (2018). elmNNRcpp: The Extreme Learning Machine Algorithm. R package version 1.0.1. https://CRAN.R-project.org/package=elmNNRcpp

Orr, R. J., Murray, P. J., Eyles, C. J., Blackwell, M. S. A., Cardenas, L. M., Collins, A. L. et al. (2016). The North Wyke Farm Platform: effect of temperate grassland farming systems on soil

moisture contents, runoff and associated water quality dynamics, European Journal of Soil Science, 67, 374–385.

Payne, R. W., Baird, D. B., Cherry, M., Gilmour, A. R., Harding, S. A., Lane, P. W., Morgan, G. W. et al. (2002). GenStat Release 6.1 Reference Manual. Part 2. Directives. Hemel Hempstead: VSN International. https://repository.rothamsted.ac.uk/item/88z4y/genstat-release-6-1-reference-manual-part-2-directives.

Percival, D. B. and Guttorp, P. (1994). Long-Memory Processes, the Allan Variance and Wavelets, Wavelet Analysis and Its Applications, 4, 325-244. Wavelets in Geophysics. Academic Press. doi: 10.1016/B978-0-08-052087-2.50018-9.

Percival, D. B. and Walden, A. T. (2000). Wavelet Methods for Time Series Analysis, Cambridge University Press.

Rust, W., Corstanje, R., Holman, I. P. and Milne, A. E. (2014). Detecting Land Use and Land Management Influences on Catchment Hydrology by Modelling and Wavelets, Journal of Hydrology, 517, 378-389. doi: 10.1016/j.jhydrol.2014.05.052.

San Martín, C., Milne, A. E., Webster, R., Storkey, J., Andújar, D., Fernández-Quintanilla, C., and Dorado, J. (2018). Spatial Analysis of Digital Imagery of Weeds in a Maize Crop, ISPRS International Journal of Geo-Information, 7(2), 61. doi: 10.3390/ijgi7020061.

Scarrott, C. and MacDonald, A. (2012). A Review of Extreme Value Threshold Es-Timation and Uncertainty Quantification, REVSTAT–Statistical Journal, 10(1), 33-60.

Smith, P., Smith, J. U., Powlson, D. S., McGill, W. B., Arah, J. R. M., Chertov, O. G. Coleman, K. et al. (1997). A Comparison of the Performance of Nine Soil Organic Matter Models Using

Datasets from Seven Long-Term Experiments, Geoderma, Evaluation and Comparison of Soil Organic Matter Models, 81(1), 153-225. doi: 10.1016/S0016-7061(97)00087-6.

Southworth, H., Heffernan J. E. and Metcalfe, P. D. (2018). texmex: Statistical modelling of extreme values. R package version 2.4.2.

Sun, Z. L., Choi, T. M., Au, K. F. and Yu, Y. (2008). Sales Forecasting Using Extreme Learning Machine with Applications in Fashion Retailing, Decision Support Systems, 46(1), 411-419. doi: 10.1016/j.dss.2008.07.009.

Takahashi, T., Harris, P. M., Blackwell, S. A., Cardenas, L. M., Collins, A. L., Dungait, J. A. J., Hawkins, J. M. B. et al. (2018). Roles of Instrumented Farm-Scale Trials in Trade-off Assessments of Pasture-Based Ruminant Production Systems, Animal, 12(8),1766-1776. doi: 10.1017/S1751731118000502.

Tsutsumida, N., Rodríguez-Veiga, P., Harris, P., Balzter, H. and Comber, A. (2019). Investigating Spatial Error Structures in Continuous Raster Data, International Journal of Applied Earth Observation and Geoinformation, 74, 259-268. doi: 10.1016/j.jag.2018.09.020.

Webster, R., and Oliver, M. A. (2007). Geostatistics for Environmental Scientists, John Wiley & Sons.

Wu, L., McGechan, M. B. McRoberts, N., Baddeley, J. A. and Watson, C. A. (2007). SPACSYS: Integration of a 3D Root Architecture Component to Carbon, Nitrogen and Water Cycling-Model Description, Ecological Modelling, 200(3), 343-359. doi: 10.1016/j.ecolmodel.2006.08.010.

Yaseen, Z. M., Sulaiman, S. O., Deo, R. C. and Chau, K. W. (2019). An Enhanced Extreme Learning Machine Model for River Flow Forecasting: State-of-the-Art, Practical Applications in Water Resource Engineering Area and Future Research Direction, Journal of Hydrology, 569, 387-408. doi: 10.1016/j.jhydrol.2018.11.069.

6. General Discussion and Conclusions

This discussion draws together the interpretations of the preceding chapters and is presented in the broader context of the aims and objectives of the thesis. The first section discusses the findings in terms of what is known from the literature and how this research adds to it through the four research questions set out in the thesis introduction. Section two presents limitations of this research, which lays the groundwork for the third section, where the implications of the research and recommendations for further work are discussed. The fourth and final section concludes the study.

6.1 Key findings from this research

6.1.1 Research question 1: Statistical modelling of extremes, threshold selection and scale effects

Modelling and forecasting of water flow, especially the peaks, is a challenging task, whether statistical, process-based or hybrid models are used. From a statistical point of view, modelling the tail of the empirical distribution of a flow time series by the Generalized Pareto Distribution (GPD) model poses several challenges. The performance of the model depends on several factors, which can introduce a high degree of uncertainty. The findings of Chapter 2 accord with previous studies which showed that the performance of the estimators of GPD parameters depends greatly on sample size and the shape characteristics of the distribution. For this reason, and within the same application, different estimators could be used according to their strengths and the characteristics of the data. For example, one estimator might be more suitable for a large sample size and positive shape parameters and another estimator where the sample size is reduced or the shape characteristics are negative. This issue is driven

by threshold selection, as the sample size is a function of the length of the time series and the threshold above which the variable is considered extreme. At the same time, the estimated parameters can affect the selected threshold when used to define it. All such issues must be considered carefully when fitting the GPD, as they can affect significantly any extrapolations beyond the observed data, which are commonly required (Hawkes et al., 2002; Liang et al., 2019).

In practical cases, the threshold can be estimated based on existing information (i.e. historical data) or based on a direct physical meaning (e.g. the threshold at which a flood is deemed a threat). However, this is not always possible, due mainly to poor data availability. When thresholds for tens or hundreds of catchments need to be estimated, threshold selection has to be done automatically and with reduced subjectivity. The technique proposed in Chapter 2 (and published in Curceac et al., 2020b) is based on threshold stability and automates a commonly used graphical method. This approach provided more robust results when compared to other explored (analytical) techniques as it was less sensitive to the shape characteristics, sample size and data scale. For this method, splines were used to locate a relatively stable region of shape and modified scale parameters. This new method shows great promise, and has already drawn attention (Willkofer et al., 2020). There is scope to develop further this graphical methodology, for example replacing splines with alternative smoothers (Härdle, 1990).

6.1.2 Research question 2: PBM simulation of peak flows and the importance of process scale

Accurate modelling and forecasting of water runoff from agricultural land is important for management of nutrient losses and water pollution. In the context of grassland agriculture,

water flow is most commonly modelled using process-based models (PBMs) (Chanasyk et al., 2003; Leitinger et al., 2010). Agricultural land is typically managed at the field-scale, where PBMs are the mathematical representation of various processes within the field. Some of these processes can be estimated and integrated in the model in a straightforward way. Other processes, for example, soil compaction caused by livestock or machinery, are more difficult to parametrise and can be a significant source of uncertainty. The parameters of the processes that are incorporated in a PBM are, in general, calibrated to capture the general trends or the means of these processes. In other words, they are structured to replicate well the central part of the distribution of a variable and for this reason they often fail in capturing the extreme events which are in the tails of the distribution. In this research (Chapters 3, 4 and 5), the SPACSYS PBM of Wu et al., (2007) was used. This model is a widely-used and well-proven model which captures accurately many of the processes central to grassland agriculture, including hydrological processes (Chanasyk et al., 2003; Leitinger et al., 2010; Pol-van-Dasselaar, 2015; Perego et al., 2016; Alaoui et al., 2018; Adimassu et al., 2019). As such, it can be considered as representative of many PBMs found in the literature.

The SPACSYS model was developed originally to run on a daily time-step. At this resolution it exhibits the common issues of underestimating the peak flows. The research presented in Chapter 3 confirmed that simulating at finer resolutions (specifically 15-minutes and hourly) and then aggregating to the daily scale can increase the accuracy compared to simulations run at a coarse daily resolution to begin with. Despite this being a relatively well known phenomenon in the field of hydrology and engineering, it was a more novel finding in the context of grassland agriculture. Reducing the time step of the model's simulation and input data has been shown to represent more accurately the infiltration-runoff partitioning

(Regenass et al., 2021), which could be a contributing factor to SPACSYS improved performance. More pertinently, this research (Chapter 3) showed that aggregating finer scale simulations (15-minute and hourly) results in peaks being more accurately identified as the smoothing effect that characterises coarse resolution simulated data is not "present". However, finer scale simulations can also result in overestimated magnitude and number of extremes compared to the measured ones. Interestingly, the statistical analysis of peak flows at different scales modelled and simulated by a GPD, also showed the greatest agreement at the hourly resolution, but for a different sub-catchment of the NWFP (Chapter 2). A tentative explanation is that the hourly data is close to the natural water run-off integration rate to each sub-catchment's water flume following a rainfall event.

6.1.3 research question 3: Integrating statistically-based models of extremes with a PBM to improve the prediction of peak flows

Hybrid modelling is the combination of two or more methods in a single modelling framework. It can include models from the same or similar disciplines (e.g. combined statistical and machine learning methods) or models that come from the physical and statistical domains. Statistical post-processing of physical model simulations is probably the most common way to combine PBM and data-driven methods and this research can be considered as a contribution in this direction (Cloke and Pappenberger, 2009; Bradley et al., 2015; Li et al., 2017). The novelty of the hybrid approach proposed in this research (Chapter 4) is the type of models chosen for this and how they were integrated together.

The conditional extreme model (CEM) is a regression-type model but with a considerably more complicated structure than the commonly used ones (e.g. least squares). Prior to this,

the statistical multivariate models for extremes required that the extremes in all the variables occurred simultaneously and had similar shape characteristics and, thus, the CEM was developed to overcome these constraints. The CEM allows for greater flexibility and has been used mainly to model the same variable at different locations. e.g. waves (Jonathan et al., 2013) or different variables measured at one site, e.g. various air quality variables (Heffernan and Tawn, 2004). In this research, the CEM was applied in an original way and used to link the same underlying variable captured by different sources (i.e. measured flow and PBM simulated flow). An important issue regarding the application of the CEM is to ensure that all the independence assumptions are respected. Similar to other stochastic models, the CEM allows the generation of a large number of random simulations defined by the user and preserves some specific stochastic characteristics of the variable of interest, which in this case is the dependence structure of the extremes. In the research presented in Chapter 4, 50 000 random simulations were obtained and their median value was calculated but the same stochastic simulations can be used to create confidence intervals that can express the uncertainty in the obtained forecasts.

The ELM model was also applied in a novel context. To the best of my knowledge, its application in a post-processing framework of PBM simulations and the hybridisation with a PBM and other statistical models has not been attempted before. Previous applications of ELM for flow forecasting that can be found in the literature use as inputs weather data, e.g. precipitation and temperature (Deo and Şahin, 2016) or measured flow at lag times (Yaseen et al., 2016). Weather variables are used as inputs to SPACSYS and therefore could not be used again in the ELM model. However I explored introducing soil moisture as a covariate in

the ELM model reported in Chapter 5 as it integrates the effects of precipitation and evapotranspiration.

Chapter 4 showed that the accuracy of the daily simulated peaks can be increased by hybridising the PBM with the CEM and ELM. Results reported in Chapter 3 showed that simulating at hourly resolution and aggregating to daily resolution improves PBM performance, including better capturing of the extremes. The combination of these approaches, namely the application of PBM-CEM-ELM hybrid on the hourly aggregated to daily simulated flow, along with the use of soil moisture was expected to improve the results further and this assessment was performed in Chapter 5.

6.1.4 Research question 4: Alternative characterisations of model performance through variography and wavelet analyses

The performance of the modelling approaches described through Chapters 2, 3 and 4 was evaluated by commonly used error and agreement indices, where measured flow was directly compared to the predicted flow. Indices including MSE, PBIAS, NSE and KGE were calculated so that models could be objectively evaluated on a point prediction basis. Each of these indices is useful to describe specific aspects of the performance of a model. For example, MSE incorporates variance and bias while KGE includes the correlation, the ratio of the means and the variability ratio. In general, error indices describe how close the total volume of the modelled hydrograph is to the measured one. The agreement indices show how accurately the dynamic of the processes is captured by the model and are sensitive to extreme values. Therefore, using more than one index is recommended as it provides a more holistic evaluation of model performance but again within specific limits. A comparison of NSE values

reported in the literature shows that SPACSYS simulates the flow as accurately as other commonly used hydrological models, such as the Soil and Water Assessment Tool (SWAT). Values of 0.56 (Du et al., 2009), 0.55 (Ikenberry et al., 2017), 0.65 (Duru et al., 2018) have been reported for daily flow simulations, while it is usual to obtain higher NSE values when flow is simulated at monthly scale (Ikenberry et al. 2017; Duru et al., 2018) or using data driven approaches (Rezaie-Balf et al., 2019).

The model performance tools used in Chapter 5 (variography and wavelets analyses) provided a different perspective and gave new insights to a given model's performance, in terms of how the models capture the observed patterns in (co-)variation of the measured flow process across temporal scales. Variograms provided a global and broad assessment of the structural temporal dependence in the measured and modelled flow data and could identify periodicities. On the other hand, wavelets give a more local and detailed evaluation of variance in the measured and modelled data across scales. Wavelets are able to identify changes in variation for a specific scale over time, as well as scale dependencies in flow behaviour. Furthermore and useful in the context of extremes, a wavelet change point analysis on measured and predicted flow can be used to better understand the nature of peak flow processes. In Chapter 5 it was reported that the magnitude of local wavelet variance was related to the likely number of peak flow events and how soon a peak flow event is likely to occur.

Wavelets not only provide a useful diagnostic tool for hydrological processes but have been used in combination with ANNs, similar to the one used in this study, to create prediction models (see Honorato et al., 2018). The flow data is initially pre-processed and decomposed into signals, which are then used as inputs in the neural network models. This technique has

been shown to provide more accurate forecasts than using the original flow time series at lag times as inputs, especially in cases when the data is characterised by high non-linearity and non-stationarity. Makwana and Tiwari (2014) also showed that the wavelets-ANN hybrid modelling can simulate flows outside the range of values used for training and extreme flows, depending on the wavelet function.

Interestingly, the inclusion of soil moisture as a predictor did not improve the results. This could be attributed to the fact that the soil moisture is already included as an internal state variable in SPACSYS. The previous study on the Farm Platform during 2011 and 2013 (Wu et al., 2016) has shown that SPACSYS most commonly overestimates the soil water content during wet periods, which could explain the increased wavelet variance at the finer scale interval. This behavior could be attributed to the phenomenon of equifinality, with the models parameters being calibrated to simulate accurately some processes, such as water fluxes, and predict other processes, such as soil moisture and water chemistry variables, with less accuracy. Other possible explanations are that the seasonality in the soil moisture is already captured by SPACSYS or the mismatch in the spatial variability as discussed in Section 5.5. Chapter 5 provides a natural stopping point for this research as it brings together key insights, models and advances presented in Chapters 2, 3 and 4.

6.2 Limitations

The main limitation of this research is that all the applied and proposed methods were tested only on data from the NWFP. In general, models need to be tested for a range of large datasets with different characteristics before being applied in practice. However, this does not make this thesis case study oriented. SPACSYS can be considered as representative of many PBMs

for simulating hydrological and other processes within a grassland field. For example, the issue of underestimating peak flows and overestimating the flows preceding and following peaks is commonly encountered in hydrological models (Newman et al., 2015; McMillan et al., 2016). The proposed hybrid methodology is tailored to this problem, as usually statistical methods are driven by data issues. Therefore, it could be assessed using other physical-based models and hydrological monitoring schemes without loss of generality.

The use of the NWFP as a case study is also novel as it uniquely provides data measurements at fine temporal and (in context with other monitoring sites) spatial resolution for multiple variables for a grassland site. I focused only on water flow because that was considered the most reliable long-term variable in the time series available from the NWFP. All the proposed techniques, and especially the hybrid model, would be suitable for other NWFP data with structured dependency of extremes (e.g. water temperature, precipitation). It is expected that these models would be inappropriate for variables that exhibit highly irregular outlier-type extremes (e.g. greenhouse gas emissions, which are also measured on the NWFP) or variables that exhibit a monotonous behaviour or have a binary effect, as for the soil moisture variable used in Chapter 5. Furthermore, the NWFP provides data measurements for the same variable from all 15 of its sub-catchments. Here, I used flow measured from two sub-catchments only. This was due to data quality issues at the outset of my study (which have since been largely addressed by the platform curators). Specifically, the research in Chapter 2 focused on flow data from sub-catchment 3 as this catchment had fewest missing values. Subsequently, we focused on sub-catchment 6 because of the readily available model (PBM) input data for this sub-catchment.

6.3 Future research

Given the methods explored and developed in this research were tested only on the NWFP, future research should assess these methods to catchments with different physical properties and weather patterns. For example, large catchments can have lag times of hours between a heavy precipitation event and peak discharge due to the distance the water has to travel, and vegetation and soil moisture conditions. The performance of the hydrological models is known to vary across seasons and catchment characteristics. The issues of underprediction of extremes in SPACSYS are encountered commonly and our proposed hybrid technique was adapted to them. It is likely that the techniques that constitute the hybrid approach would need to be re-calibrated in cases where a different behaviour or performance was observed. It would also be intriguing to assess the hybrid approach on a purely hydrological model and for catchments very different from the one used in this research. If modelled at various temporal resolutions, the comparison between the aggregated fine and coarser resolution simulations could provide new insights into the process. These issues could also impact the performance of the proposed threshold selection technique since processes which contribute to the generation of flow and are likely to behave differently compared to the ones used here, would normally result in different shape and scale characteristics. Testing the proposed approaches on larger datasets with considerably longer periods of record would reduce the uncertainty (e.g. more data to train the models) and allow for greater flexibility (e.g. stricter peaks independence criteria or compare Peaks over Threshold with the block maxima).

This research could be further expanded to other application areas. The proposed approaches could be transferred readily and tested on other variables such as soil and water chemistry. Hybrid models would have to be trained and calibrated to account for the different

characteristics of these simulated processes. Most of the temporal and spatial datasets measured at the NWFP provide numerous possibilities to further explore patterns, characteristics and so improve modelling and the prediction of extremes of other important processes. For example, the development of models that use the fine temporal autocorrelation of a variable (e.g. nutrient flows) at one sub-catchment, or use the correlation of the same variable measured at all 15 sub-catchments or use the spatial and temporal correlation of different variables across all 15 sub-catchments. Such models can extend those presented here or be newly developed in an exploratory context or integrated in a different inference framework, for example Bayesian (Hsu et al., 2009; Dotto et al., 2011). Improved model performance could also be obtained by implicitly accounting for spatial autocorrelation via a distance- or contiguity-based spatial weighting operation. The development and implementation of such models could also be used in tandem to help solve data quality control issues such as the imputation of missing values and outlier detection. The improved simulations of waterflow can also be used as a covariate to increase the accuracy when simulating other, more pertinent processes, such as greenhouse gas emissions and nutrient losses.

In this research, I explored new and existing models of water flow in the context of peak flow events. Typical of most models that describe environmental systems, these models are subject to uncertainties both in their structure and parameterisation. This issue was explored in Chapter 2 where the uncertainty in the estimates of the GPD parameters was assessed by using different estimators. In classical frequency analysis, the parameters are assumed to be independent and identically distributed. However, this assumption is frequently not met as the hydro-climatic extremes have been shown to be influenced by large-scale low frequency

climate variation (e.g. Enfield et al., 2001; Ouachani et al., 2013; Ouarda et al., 2014). Over the past centuries, floods in Britain seem to be related to negative phases of the North Atlantic Oscillation (NAO) (Macklin and Rumsby, 2007). The NAO has also been shown to be positively correlated with winter runoff in the south-west of England (Shepherd et al., 2017). The negative phases of the monthly NAO during 2012 and 2013 and at the end of 2015 (Chazette, 2020) could be associated with the highest peak flow events observed in this study. The shape parameter of the GPD is usually assumed constant and stationary but there is evidence that climate oscillations affect the shape of the distribution (Ouarda and Charron, 2019). For future research, such non-stationary effects can be modelled by allowing the model's parameters to be conditional on covariates representing climate variation.

Most PBMs are designed at a fixed temporal resolution. Due to the increasing data availability from different sources and higher sampling rates, there is a need for PBMs that allow greater flexibility for the end-user to choose the required simulation resolution according to the level of detail required for a specific application. However, PBM parameters are usually best calibrated for a specific temporal resolution and applying the same equations at different time scales can be a difficult task (Blöschl and Sivapalan, 1995). In cases when the PBM outputs are at a coarser scale than the one required, downscaling can be applied.

The research described in this thesis has practical value to policy makers and environmental regulators who are interested in the prediction of peak flow events both in the short- and medium-time frame. Given an accurate weather forecast, a PBM can be used to predict flow, for example, with a time frame of a week. This also holds true for hybrid models, given sufficient historic data for training. Any covariates used in the models would also need to be independently predicted into the future. Predictions over a longer-time frame will be less

accurate as they rely on weather forecasting. However, characterising the long-term behaviour of a catchment (either statistically or through hybrid modelling) may be useful for risk analysis as it should enable planning for mitigating risks and fairer insurance estimates. For these types of models to be taken up more widely in policy and regulation, further development is needed in improving data collection and quality assurance.

6.4 Concluding remarks

Accurate modelling and forecasting of water flow is of great importance but also a challenging task. Hence, researchers have developed a plethora of methods that aim to replicate characteristics such as the general trend, the mean, the variance and the correlation structure. Peak flows are most commonly underestimated. This thesis explored various techniques with which to more accurately represent peak flow events at varying scales, and proposed new ones.

The statistical analysis of extremes by the GPD showed that careful selection of the distribution parameter estimators and of the threshold that defines the extremes are crucial. When thresholds cannot be chosen according to physically meaningful values, it is important that they are estimated by reproducible methods that don't rely on subjective and biased visual interpretation. The methods proposed in this research fulfill this criterion. They demonstrated the advantage of using the most accurate and unbiased estimators for robust and automated threshold estimates, which is essential if methods are to be upscaled for practical use across a large number of applications.

Reliable forecasts of extreme flow are essential for flood protection and the improved management of grassland systems. The commonly-used daily time step of PBM simulations

results in a smoothing effect that underestimates the maximum flow and overestimates the flows preceding and following it. Simulating at finer resolutions and then upscaling to the daily scale can significantly improve the identification of the number of peak events and increase the accuracy of their forecasted magnitude. Combining statistical and data driven models (such as CEM and ELM) with PBM outputs in a post-processing framework offers similar improvements. The combining fine-resolution PBM outputs with the characteristics of CEM and ELM results in more accurate representations of the dynamics of the flow process, and therefore, increases the accuracy of the forecasts.

The performance of models is often assessed with metrics that focus on mean and variance. When the focus of prediction is on extreme values or some other quantile in the distribution, it is particularly important to consider various methods for model evaluation. Indices such as MSE, PBIAS, NSE and KGE are useful to evaluate on a point prediction basis, with each describing specific aspects of the performance of a model. In the case of time series prediction, variography and wavelet analysis complement these metrics. These methods provided unique insights into each models' ability to capture the (co)-variation in the signal across temporal-scales, in terms of both the total hydrograph and peak events. In particular the non-stationary nature of the wavelet transform allows the analyst to determine whether the model is picking up significant changes in variation at each scale-interval. This is particularly useful when analysing time series where extreme events and bursts of activity are anticipated. In future, it would be of great interest to combine the point prediction of the hybrid models with the pattern matching diagnostics in an integrated optimization approach.

References

Adimassu, Z., G. Alemu and L. Tamene. (2019). Effects of Tillage and Crop Residue Management on Runoff, Soil Loss and Crop Yield in the Humid Highlands of Ethiopia, Agricultural Systems, 168, 11-18. https://doi.org/10.1016/j.agsy.2018.10.007.

Alaoui, A., M. Rogger, S. Peth, and G. Blöschl. (2018). Does Soil Compaction Increase Floods? A Review. Journal of Hydrology, 557, 631-642. https://doi.org/10.1016/j.jhydrol.2017.12.052.

Blöschl, G., and M. Sivapalan. (1995). Scale Issues in Hydrological Modelling: A Review. Hydrological Processes, 9(3–4), 251-290. https://doi.org/10.1002/hyp.3360090305.

Bradley, A. A., M. Habib, and S. S. Schwartz. (2015). Climate Index Weighting of Ensemble Streamflow Forecasts Using a Simple Bayesian Approach. Water Resources Research, 51(9), 7382-7400. https://doi.org/10.1002/2014WR016811.

Chanasyk, D. S, E Mapfumo, and W Willms. (2003). Quantification and Simulation of Surface Runoff from Fescue Grassland Watersheds. Agricultural Water Management, 59(2), 137-153. https://doi.org/10.1016/S0378-3774(02)00124-5.

Chazette, P. (2020). Aerosol optical properties as observed from an ultralight aircraft over the Strait of Gibraltar. Atmos. Meas. Tech., 13, 4461-4477. https://doi.org/10.5194/amt-13-4461-2020.

Cloke, H. L., and F. Pappenberger. (2009). Ensemble Flood Forecasting: A Review. Journal of Hydrology, 375(3), 613-626. https://doi.org/10.1016/j.jhydrol.2009.06.005.

Curceac, S., P. M. Atkinson, A. Milne, L. Wu, and P. Harris. (2020). An Evaluation of Automated GPD Threshold Selection Methods for Hydrological Extremes across Different Scales. Journal of Hydrology, 585, 124845. https://doi.org/10.1016/j.jhydrol.2020.124845.

Deo, R. C, and M. Şahin. (2016). An Extreme Learning Machine Model for the Simulation of Monthly Mean Streamflow Water Level in Eastern Queensland. Environmental Monitoring and Assessment, 188(2), 90. https://doi.org/10.1007/s10661-016-5094-9.

Dotto, C. B. S., M. Kleidorfer, A. Deletic, W. Rauch, D. T. McCarthy, and T. D. Fletcher. (2011). Performance and Sensitivity Analysis of Stormwater Models Using a Bayesian Approach and Long-Term High Resolution Data. Environmental Modelling & Software, 26(10), 1225-1239. https://doi.org/10.1016/j.envsoft.2011.03.013.

Du, B., Xiaoyi J., R. D. Harmel and L. M. Hauck. (2009). Evaluation of a Watershed Model for Estimating Daily Flow Using Limited Flow Measurements. JAWRA Journal of the American Water Resources Association, 45(2), 475-484. https://doi.org/10.1111/j.1752-1688.2009.00303.xDuru U., Arabi M., Wohl E. E. (2018). Modeling stream flow and sediment yield using the SWAT model: a case study of Ankara River Basin, Turkey, Phys Geogr, 39(3), 264–289. doi:10.1080/02723646.2017.1342199.

Enfield, D. B., A. M. Mestas-Nuñez, and P. J. Trimble. (2001). The Atlantic Multidecadal Oscillation and Its Relation to Rainfall and River Flows in the Continental U.S. Geophysical Research Letters, 28(10), 2077-2080. https://doi.org/10.1029/2000GL012745.

Härdle, W. (1990). Applied Nonparametric Regression. Econometric Society Monographs. Cambridge: Cambridge University Press. https://doi.org/10.1017/CCOL0521382483.

Hawkes, P. J., B. P. Gouldby, J. A. Tawn, and M. W. Owen. (2002). The Joint Probability of Waves and Water Levels in Coastal Engineering Design. Journal of Hydraulic Research, 40(3), 241-251. https://doi.org/10.1080/00221680209499940.

Heffernan, J. E., and J. A. Tawn. (2004). A Conditional Approach for Multivariate Extreme Values (with Discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(3), 497-546. https://doi.org/10.1111/j.1467-9868.2004.02050.x.

Honorato, A.G., Silva, G.B., and Santos, C.A. (2018). Monthly streamflow forecasting using neuro-wavelet techniques and input analysis. Hydrological Science Journal, 63 (15–16), 2060-2075. https://doi.org/10.1080/02626667.2018.1552788

Hsu, K., H. Moradkhani, and S. Sorooshian. (2009). A Sequential Bayesian Approach for Hydrologic Model Selection and Prediction. Water Resources Research, 45(12). https://doi.org/10.1029/2008WR006824.

Ikenberry, C. D., M. L. Soupir, M. J. Helmers, W. G. Crumpton, J. G. Arnold, P. W. Gassman. (2017). Simulation of Daily Flow Pathways, Tile-Drain Nitrate Concentrations, and Soil-Nitrogen Dynamics Using SWAT, JAWRA Journal of the American Water Resources Association, 53(6), 1251-1266. https://doi.org/10.1111/1752-1688.12569.

Jonathan, P., K. Ewans, and D. Randell. (2013). Joint Modelling of Extreme Ocean Environments Incorporating Covariate Effects. Coastal Engineering, 79, 22-31. https://doi.org/10.1016/j.coastaleng.2013.04.005.

Leitinger, G., E. Tasser, C. Newesely, N. Obojes, and U. Tappeiner. (2010). Seasonal Dynamics of Surface Runoff in Mountain Grassland Ecosystems Differing in Land Use. Journal of Hydrology, 385(1), 95-104. https://doi.org/10.1016/j.jhydrol.2010.02.006.

Li, W., Q. Duan, C. Miao, A. Ye, W. Gong, and Z. Di. (2017). A Review on Statistical Postprocessing Methods for Hydrometeorological Ensemble Forecasting. Wiley Interdisciplinary Reviews: Water, 4(6), e1246. https://doi.org/10.1002/wat2.1246.

Liang, B., Z. Shao, H. Li, M. Shao, and D. Lee. (2019). An Automated Threshold Selection Method Based on the Characteristic of Extrapolated Significant Wave Heights. Coastal Engineering, 144, 22-32. https://doi.org/10.1016/j.coastaleng.2018.12.001.

Macklin, M. G., and B. T. Rumsby. (2007). Changing Climate and Extreme Floods in the British Uplands. Transactions of the Institute of British Geographers, 32(2), 168–86. https://doi.org/10.1111/j.1475-5661.2007.00248.x.

Makwana J. J. and M. K. Tiwari. (2014). Intermittent Streamflow Forecasting and Extreme Event Modelling using Wavelet based Artificial Neural Networks, Water Resources Management, 28, 13, 4857-4873. https://doi.org/10.1007/s11269-014-0781-1.

McMillan, H. K., D. J. Booker, and C. Cattoën. (2016). Validation of a National Hydrological Model. Journal of Hydrology, 541, 800-815. https://doi.org/10.1016/j.jhydrol.2016.07.043.

Newman, A. J., M. P. Clark, K. Sampson, A. Wood, L. E. Hay, A. Bock, R. J. Viger, et al. (2015). Development of a Large-Sample Watershed-Scale Hydrometeorological Data Set for the Contiguous USA: Data Set Characteristics and Assessment of Regional Variability in Hydrologic

Model Performance. Hydrology and Earth System Sciences, 19(1), 209-223. https://doi.org/10.5194/hess-19-209-2015.

Ouachani, R., Z. Bargaoui, and T. Ouarda. (2013). Power of Teleconnection Patterns on Precipitation and Streamflow Variability of Upper Medjerda Basin. International Journal of Climatology, 33(1), 58-76. https://doi.org/10.1002/joc.3407.

Ouarda, T. B. M. J., C. Charron, K. N. Kumar, P. R. Marpu, H. Ghedira, A. Molini, and I. Khayal. (2014). Evolution of the Rainfall Regime in the United Arab Emirates. Journal of Hydrology, 514, 258-270. https://doi.org/10.1016/j.jhydrol.2014.04.032.

Ouarda, T. B. M. J., and C. Charron. (2019). Changes in the Distribution of Hydro-Climatic Extremes in a Non-Stationary Framework. Scientific Reports, 9(1), 1-8. https://doi.org/10.1038/s41598-019-44603-7.

Perego, A., L. Wu, G. Gerosa, A. Finco, M. Chiazzese, and S. Amaducci. (2016). Field Evaluation Combined with Modelling Analysis to Study Fertilizer and Tillage as Factors Affecting N2O Emissions: A Case Study in the Po Valley (Northern Italy). Agriculture, Ecosystems & Environment, 225, 72-85. https://doi.org/10.1016/j.agee.2016.04.003.

Pol-van-Dasselaar, A., H. F. M. Aarts, A. De Vliegher, A. Elgersma, D. Reheul, J. A. Reijneveld, J. Verloop and A. Hopkins. (2015). Grassland and Forages in High Output Dairy Farming Systems: Proceedings of the 18th Symposium of the European Grassland Federation, Wageningen, The Netherlands. Grassland Science in Europe, (20). Wageningen: Wageningen Academic Publishers.

Rezaie-Balf, M., S. Kim, H. Fallah and S. Alaghmand. (2019). Daily river flow forecasting using ensemble empirical mode decomposition based heuristic regression models: application on the perennial rivers in Iran and South Korea, J. Hydrol., 572, 470-485. https://doi.org/10.1016/j.jhydrol.2019.03.046.

Shepherd, A., W. Atuhaire, L. Wu, D. Hogan, R. Dunn, and L. Cardenas. (2017). Historic Record of Pasture Soil Water and the Influence of the North Atlantic Oscillation in South-West England. Hydrology Research, 48(1), 277-294. https://doi.org/10.2166/nh.2016.195.

Willkofer, F., R. R. Wood, F. von Trentini, J. Weismüller, B. Poschlod, and R. Ludwig. (2020). A Holistic Modelling Approach for the Estimation of Return Levels of Peak Flows in Bavaria. Water, 12(9), 2349. https://doi.org/10.3390/w12092349.

Wu, L., M. B. McGechan, N. McRoberts, J. A. Baddeley, and C. A. Watson. (2007). SPACSYS: Integration of a 3D Root Architecture Component to Carbon, Nitrogen and Water Cycling—Model Description. Ecological Modelling, 200(3), 343-359. https://doi.org/10.1016/j.ecolmodel.2006.08.010.

Yaseen, Z. M., O. Jaafar, R. C. Deo, O. Kisi, J. Adamowski, J. Quilty, and A. El-Shafie. (2016). Stream-Flow Forecasting Using Extreme Learning Machines: A Case Study in a Semi-Arid Region in Iraq. Journal of Hydrology, 542, 603-614. https://doi.org/10.1016/j.jhydrol.2016.09.035.

Appendix A. Equations of the estimators

The estimators used in this study can be formally defined as follows:

1. MLE method:

$$L = -n\log\sigma + \left(\frac{1}{\xi} - 1\right)\sum_{i=1}^{n}\log\left(1 - \frac{\xi x_i}{\sigma}\right), \quad \xi \neq 0$$

$$L = -n\log\sigma - \frac{1}{\sigma}\sum_{i=1}^{n} x_i\,, \quad \xi = 0$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are the order statistics of a random sample $x_1, \ldots, x_n$ from the GPD. The estimated parameters are obtained when the log-likelihood function L is maximized.

2. MPLE method:

$$P(\xi) = \begin{cases} 1 & \xi \leq 0 \\ \exp\{-\lambda\left(\frac{1}{1-\xi} - 1\right)^{a}\} & 0 < \xi < 1 \\ 0 & \xi \geq 1 \end{cases}$$

where a and λ are the penalizing non-negative constants. The corresponding penalized likelihood function is $L_{pen} = L \times P$.

3. LME is a combination of both likelihood and moment estimators and is derived from:

$$\frac{1}{n}\sum_{i=1}^{n}(1 - \theta x_i)^{P} - \frac{1}{1-r} = 0, \qquad \theta < x_{(n)}^{-1},$$

where $\theta = \xi/\sigma$ and $P = -\frac{rn}{\sum_{i=1}^{n}\log(1-\theta x_i)}$. The parameter $r < 1, r \neq 0$ must be pre-defined before the estimation and either be set as ξ if there is an initial estimate of it or taken as $r = -1/2$.

4. MOM estimators (Hosking & Wallis, 1987) of the scale σ and shape ξ parameters of the GPD distribution are given by:

$$\hat{\sigma} = \frac{1}{2}\bar{x}\left(\frac{\bar{x}}{s^2} + 1\right), \quad \hat{\xi} = \frac{1}{2}\left(\frac{\bar{x}^2}{s^2} - 1\right)$$

where $\bar{x}$ and s^2 are the sample mean and variance.

5. PWM estimators provide estimates with smaller bias and variance than MLE when the sample size is less than 500 (Hosking & Wallis 1987). The PWM's of the random variable *X* with a distribution function $G \equiv G(x) = P(X \leq x)$ is defined as:

$$M_{l,j,k} = E\left[X^l F^j (1-F)^k\right] = \int_0^1 [x(F)]^l F^l (1-F)^k dF$$

where l, j and k are real numbers. For $j = k = 0$ and *l* a nonnegative integer, $M_{l,0,0}$ is the classical moment of order l.

6. The estimator suggested by Pickands (1975) (referred to as 'Pick') is based on the ascending order statistics $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$ from an independent sample of size n and is defined as:

$$\hat{\xi}_{n,k}^{Pick} = \frac{1}{log2}\log\left(\frac{X_{n-k+1,n} - X_{n-2k+1,n}}{X_{n-2k+1,n} - X_{n-4k+1,n}}\right), \text{ for } k = 1, \ldots, [n/4]$$

This estimator is largely dependent on k and provides a large asymptotic variance (e.g. (Dekkers & Haan, 1989; Segers, 2005; Yun, 2002).

There are many MGF statistics that can be used for GPD parameter estimation, such as Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling (see Luceño, 2006)

Appendix B. GPD estimators performance figures

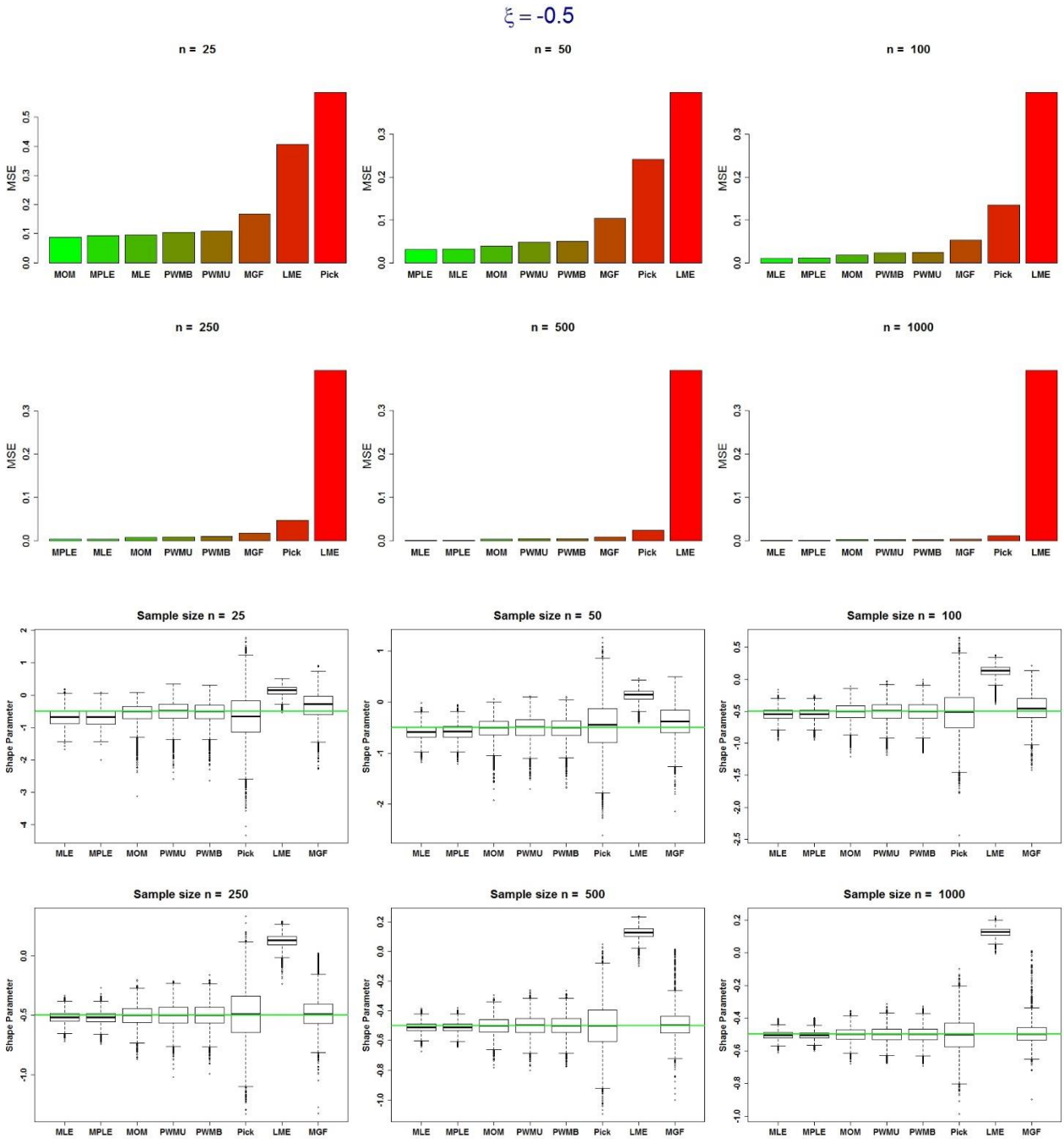


Figure B-1: Performance of GPD estimators for shape parameter $\xi = -0.5$ and for six different sample sizes ($n = 25,\ 50, 100, 250, 500, 1000$).

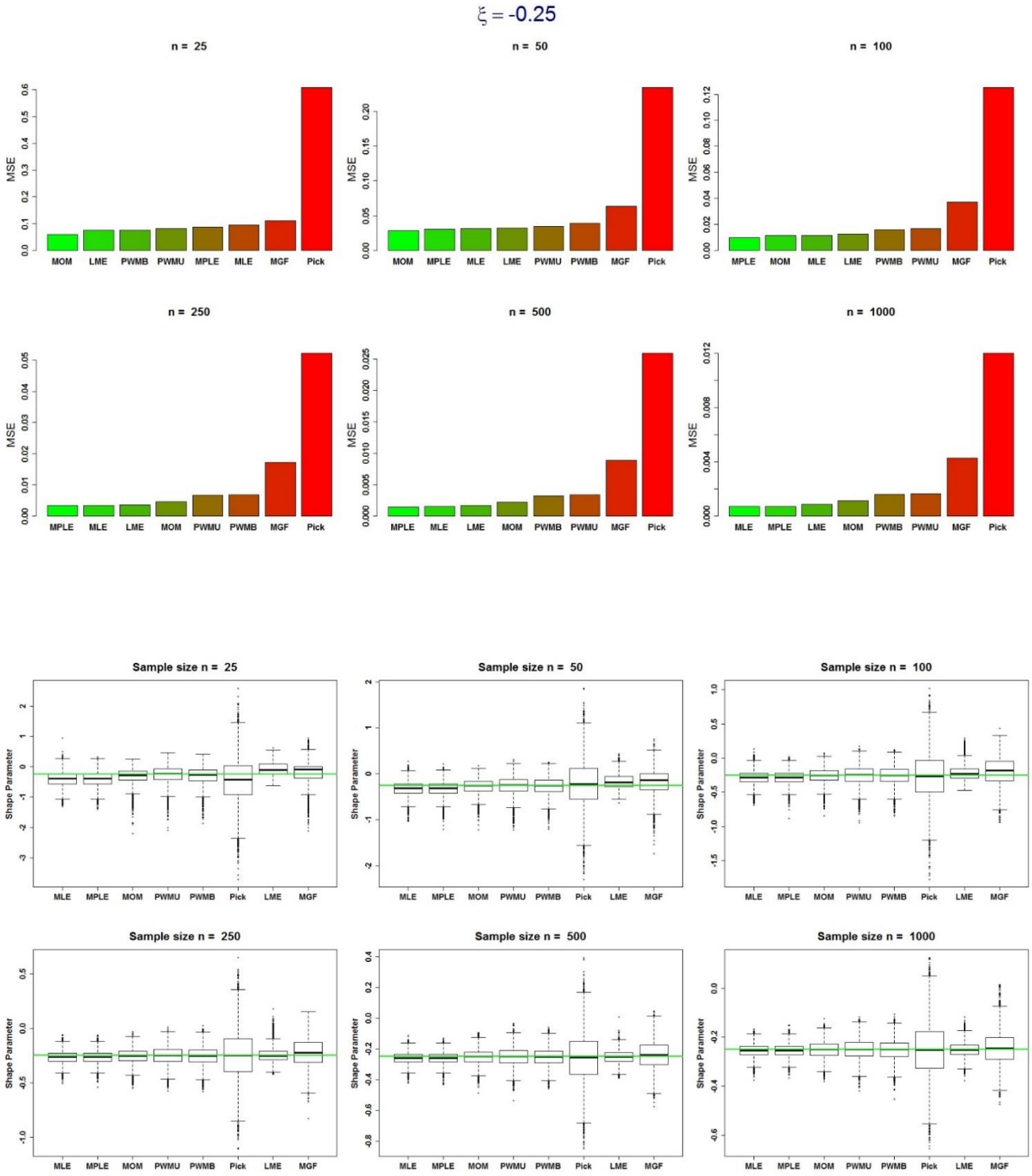


Figure B-2: Performance of GPD estimators for shape parameter $\xi = -0.25$ and for six different sample sizes ($n = 25,\ 50, 100, 250, 500, 1000$).

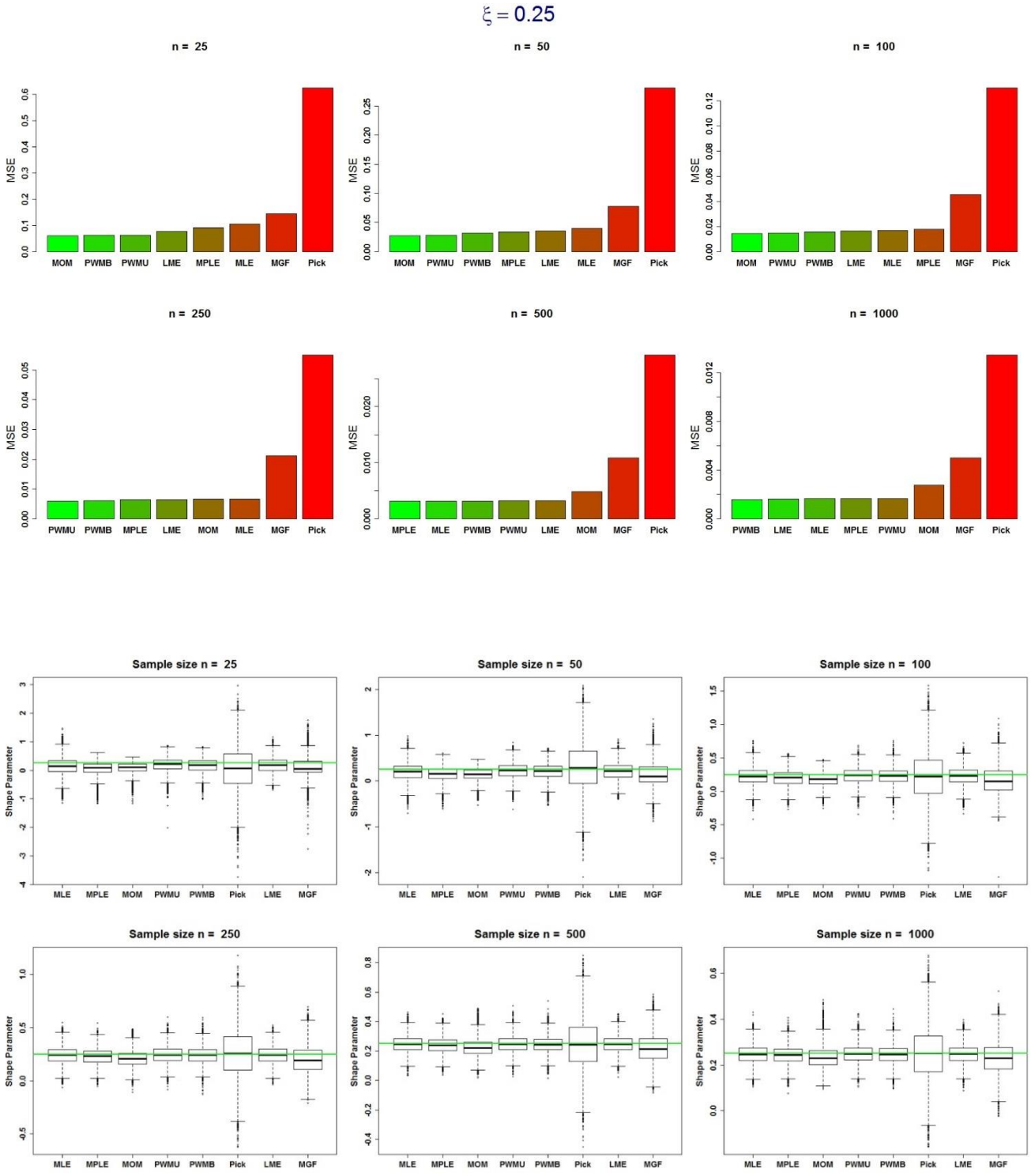


Figure B-3: Performance of GPD estimators for shape parameter $\xi = 0.25$ and for six different sample sizes ($n = 25,\ 50, 100, 250, 500, 1000$).

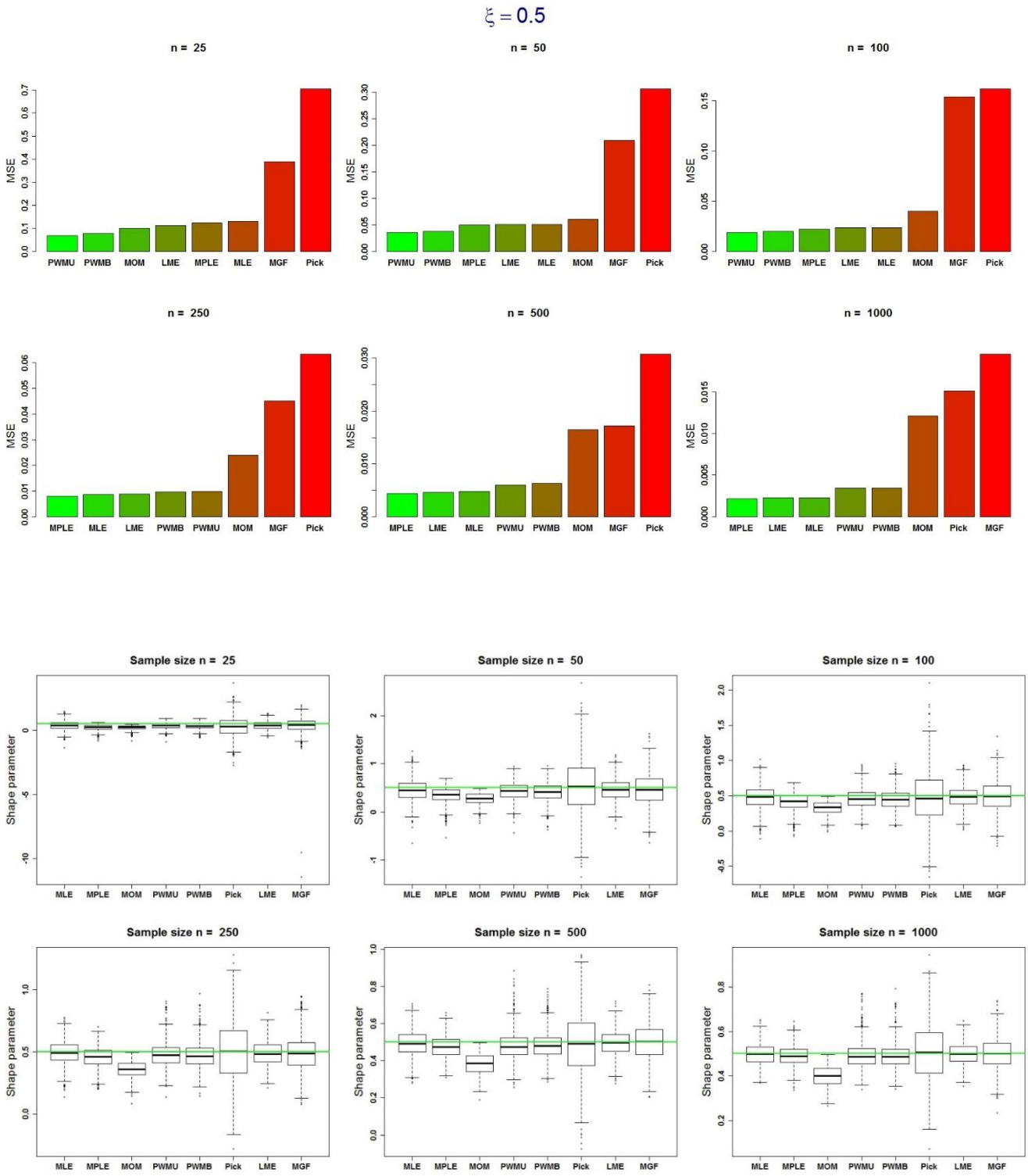


Figure B-4: Performance of GPD estimators for shape parameter $\xi = -0.25$ and for six different sample sizes ($n = 25,\ 50, 100, 250, 500, 1000$).

Appendix C. Forecasting maximum peaks by PBM, CEM and ELM

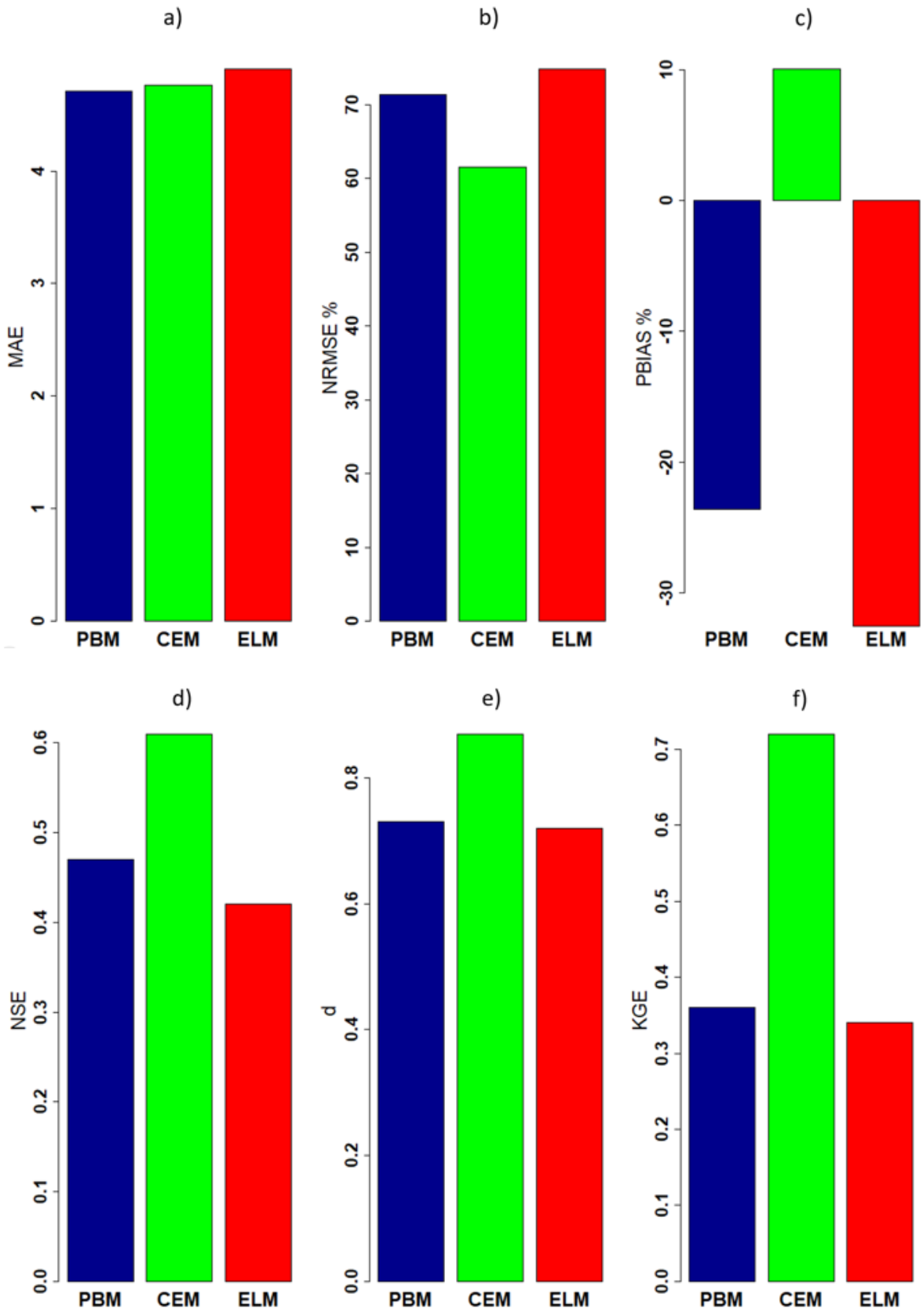


Figure C-1: Error and agreement indices of the PBM, CEM and ELM simulated maximum peaks compared to observed data. a) MAE, b) NRMSE, c) PBIAS, d) NSE, e) *d*, f) KGE.

Appendix D. Variograms models

The base models considered for variogram models were

Spherical:

$$\gamma(h) = c_0 + c\left\{\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right\} \text{ for } h \le a$$
$$= c_0 + c \text{ for } h > a$$
$$= 0 \text{ for } h = 0,$$

where h is a scalar in temporal distance only. Its parameters are c_0 which is the nugget variance, c is the correlated variance and a is the distance parameter (the range) of the model. Parameter a is the limiting distance of temporal dependence or correlation. The parameter c is the variance of the correlated structure, so that $c_0 + c$ is the total variance of the underlying random process, of which the data are a realization.

Circular:

$$\gamma(h) = c_0 + c\left\{1 - \frac{2}{\pi}\cos^{-1}\left(\frac{h}{a}\right) + \frac{2h}{\pi a}\sqrt{1 - \frac{h^2}{a^2}}\right\} \text{ for } h \le a$$
$$= c_0 + c \text{ for } h > a$$
$$= 0 \text{ for } h = 0,$$

in which the parameters c_0, c and a are defined in the same way as for the spherical model.

Exponential model:

$$\gamma(h) = c_0 + c\left\{1 - \exp\left(-\frac{h}{r}\right)\right\}$$

in which the parameters c_0 and c are defined as above but r is the distance parameter which is approximately a third of the effective range (see Webster and Oliver, 2007).

For each of our variables the double spherical proved the best model. This is given by:

$$\gamma(h) = c_0 + c_1\left\{\frac{3h}{2a_1} - \frac{1}{2}\left(\frac{h}{a_1}\right)^3\right\} + c_2\left\{\frac{3h}{2a_2} - \frac{1}{2}\left(\frac{h}{a_2}\right)^3\right\} \text{ for } h \le a_1$$

$$= c_0 + c_1 + c_2\left\{\frac{3h}{2a_2} - \frac{1}{2}\left(\frac{h}{a_2}\right)^3\right\} \text{ for } a_1 < h \le a_2$$

$$= c_0 + c_1 + c_2 \text{ for } h > a_2$$

$$= 0 \text{ for } h = 0,$$

The parameters for each model are given in the table below

Table D-1: Parameters of the variograms models

	Distance parameters		Sill parameters		Nugget
	a_1	a_2	c_1	c_1	c_0
Modelled daily	11.89	176.68	2.67	21.88	0.46
Hybrid daily	11.81	176.95	2.59	21.36	0.48
Hybrid daily with SM	11.73	176.32	2.65	21.26	0.47
Modelled H2D	10.88	204.9	4.79	8.91	5.83
Hybrid H2D	11.62	198.87	6.11	11.01	4.12
Hybrid H2D with SM	11.86	183.65	3.55	20.22	1.12
Precipitation	11.31	211	4.17	2.77	8.97