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

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 two approaches are firstly, models that aggregate fine resolution (sub-daily) simulated flow from a process-based model (PBM) to daily, and secondly, hybrid models that combine PBMs with statistical and machine learning methods. 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. 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.

1. Introduction

In many regions across the globe, changing patterns of rainfall may increase 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 flow events (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. 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 use hybrid models (Bogner et al., 2016; Papacharalampous et al., 2019). 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 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 (temporal) dependence or autocorrelation and for temporal applications is able to identify the temporal scales over which the 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 wavelet variance or wavelet correlation for a given scale interval over time

In this research, the modelling concepts described above are integrated to explore the relative increases in accuracy made 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 the conventional PBM and the aggregation-based PBM outputs. Further, we explore using 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

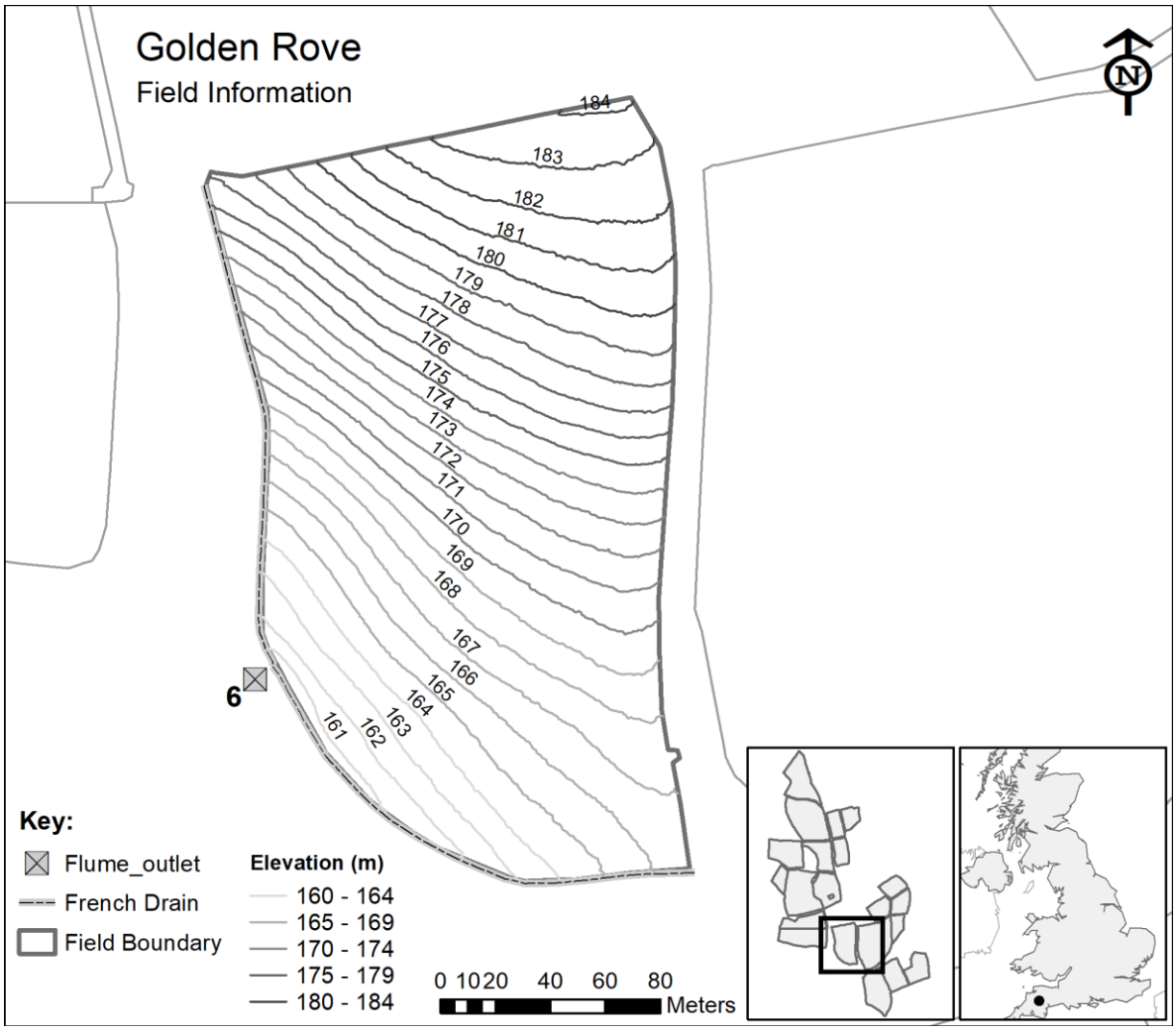
of the latter in capturing variation of the former at different scales and locations in time. Critically, we compare change points detected by the wavelet analysis (measured and modelled) with the extreme flow events suggested by the threshold selected based on stability plots of the Generalized Pareto distribution (GPD) of (Curceac et al., 2020b). The exploratory analysis 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 diagnostics 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.

2. Materials and Methods

2.1. 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 Fig. 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

115 at a 15-minute temporal frequency since October 2012 through a combination of primary and secondary
116 flow devices (as detailed in Orr et al., 2016; Curceac et al., 2020a). The flow is generated only from rainfall as
117 the fields are not irrigated. Each sub-catchment also monitors precipitation and soil moisture every 15
118 minutes at a depth of 10 cm. For this research, we used flow discharge, rainfall and soil moisture (SM) (from
119 April 2013 to February 2016) measured at sub-catchment 6 (Fig. 2), which consists of a single field (Golden
120 Rove). This field was chosen because, as part of the permanent pasture treatment of the NWFP, it would not
121 have been ploughed and reseeded during the period of study (which would affect the run-off process).



122

123 **Fig. 1.** Sub-catchment (consisting of a single field) selected from the total of 15 sub-catchments within the
124 North Wyke Farm Platform, South-West England, UK. Precipitation and soil moisture data are collected from
125 a site centrally-located in the sub-catchment.

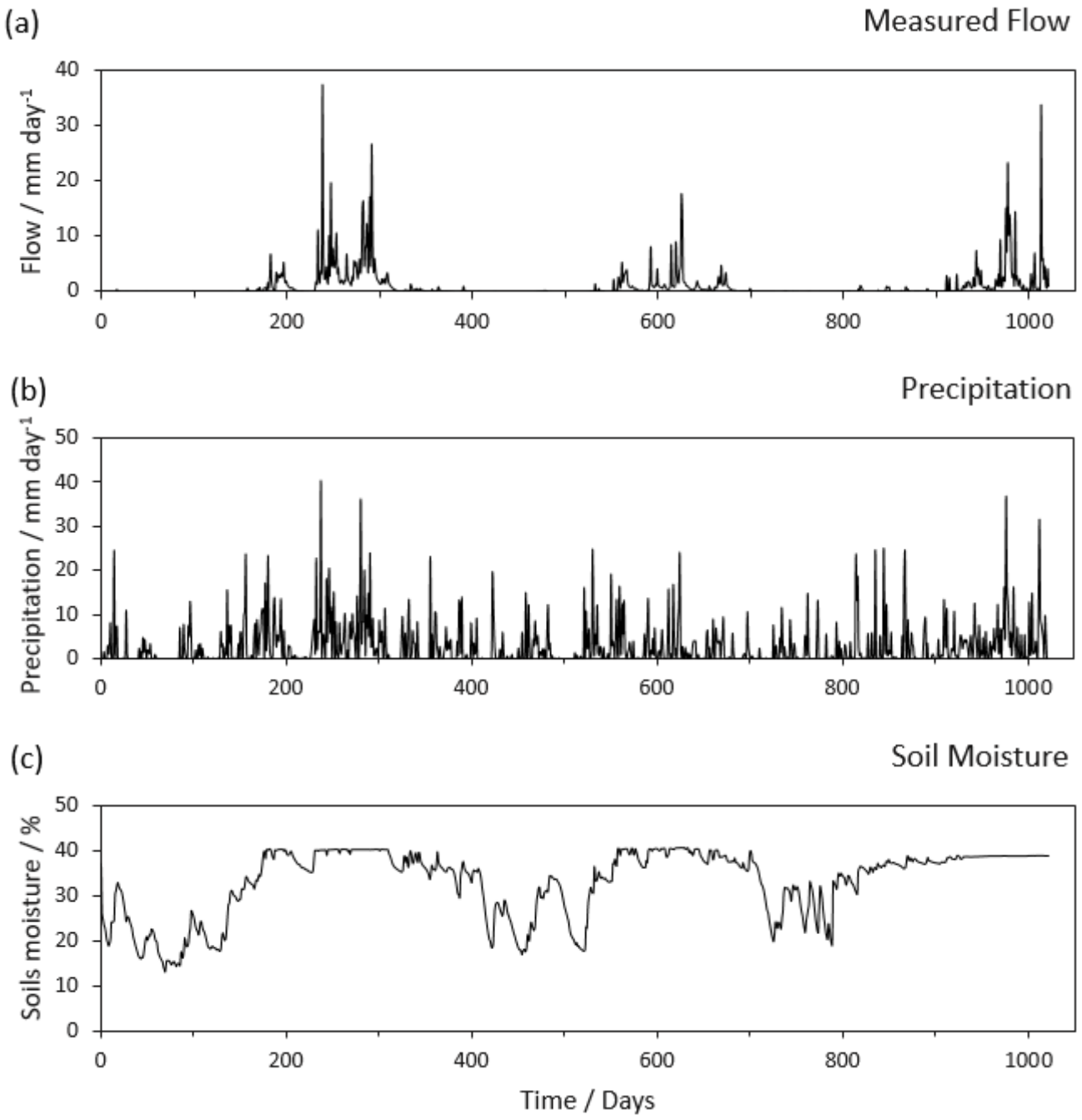


Fig. 2. (a) Flow data measured at the study site, (b) precipitation used as input in the PBM and (c) soil moisture (SM) used as a covariate in the ELM component of the hybrid model. All measurements aggregated from 15 minute to daily.

2.2. Models for simulation and forecasting

2.2.1. Process-based Model (PBM)

Flow 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 for water potential. 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.

2.2.2. 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 flow 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).

2.2.3. 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 the CEM and ELM models.

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

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

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.

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 the Laplace distribution for them to follow the same margins. The initial vector variable X is, therefore, transformed as:

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

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:

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

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:

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

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:

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

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).

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:

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

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 described later. The number of HN in the hidden layer presents a classic problem of over-fitting and under-fitting and is commonly defined empirically (Sun et al., 2008).

Both the CEM and ELM models 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, soil moisture as well. Using the data that were left out for testing purposes (except for the maximum), forecasts were obtained.

Peaks smaller than the cluster maxima were forecasted by the ELM and the CEM was used only to forecast maximum flows. 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.

2.3. Model analysis

2.3.1. Error and agreement indices

The models were evaluated initially by calculating the Pearson correlation coefficient, the mean absolute error (MAE) and the Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) indices between the measured and simulated flow. The optimal value of MAE is zero, and the smaller the value, the more accurate are the simulations. The NSE takes values from minus infinity to one, 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 simulations. The indices were calculated using the following:

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

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

where $\hat{z}_i$ are the simulated values, z_i are the measured values and $\bar{z}_i$ is the mean of the measured values.

225 **2.3.2. Variograms**

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

231 $$\gamma(h) = \frac{1}{2}\mathrm{E}[\{Z(t) - Z(t+h)\}^2],$$

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

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

235 $$\hat{\gamma}(h) = \frac{1}{2m}\sum_{i=1}^{m}[Z(t+h) - Z(t_i)]^2$$

236 where $Z(t_i)$ and $Z(t_i\ +\ h)$ are the observed values at times t_i and $t_i + h$ separated by h, and of which there
237 are $m(h)$ paired comparisons at that lag. As observations of the process become further apart (quantified by
238 h) they typically become less correlated, and often there exists a lag beyond which there is no correlation.
239 We fitted plausible models to the empirical variograms using the directive FITNONLINEAR in GenStat (v. 18)
240 (Payne et al., 2008). Authorised variogram models have simple shapes, but can be combined additively to
241 represent more complex shapes (Webster and Oliver, 2007). The base variogram models that we considered
242 were spherical, circular and exponential (see S1 for details).

243 In this research, we computed empirical and modelled variograms for measured flow, measured
244 precipitation and measured SM together with the simulated flow data from each of the six models described
245 above. For measured and modelled flow and precipitation, data were log transformed before variograms
246 were fitted because of the skew in the data (i.e., transforms were used to facilitate authorised variogram
247 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.

2.3.3. Wavelet analysis

We used the maximum overlap discrete wavelet transform (MODWT) (Percival and Walden, 2000) to analyse the performance of each model in representing scale-dependant variation in the measured flow time-series. 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 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

$$\sigma_{u,j}^2 = \frac{1}{2^j n_j} \sum_{k=1}^{n_j} \{d_{j,k}^u\}^2,$$

where $d_{j,k}^u$ is the kth 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

$$\rho_{u,v,j} = \frac{C_{u,v,j}}{\sigma_{u,j}\sigma_{v,j}}.$$

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

$$C_{u,v,j} = \frac{1}{2^j n_j} \sum_{k=1}^{n_j} d_{j,k}^u d_{j,k}^v.$$

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.

2.3.4. Change point detection with wavelets

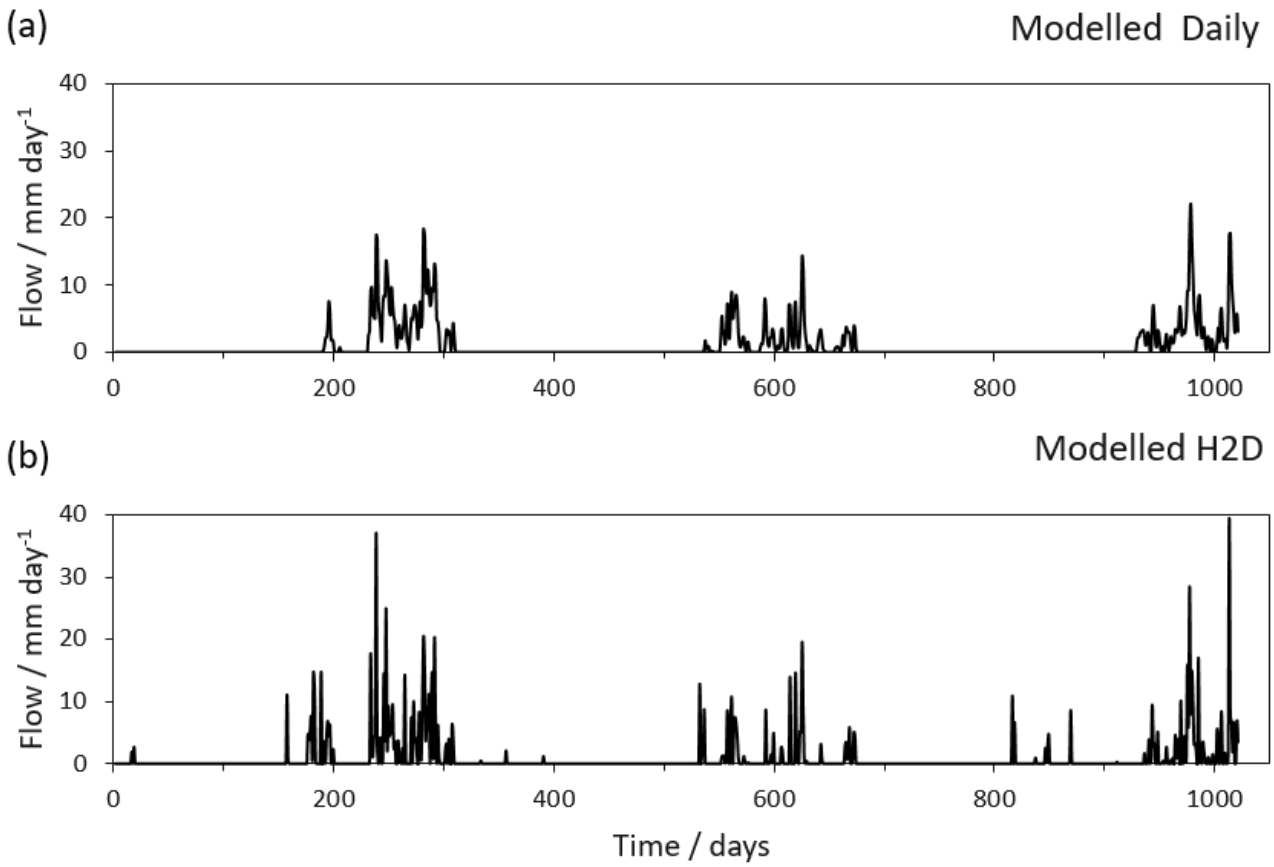
To determine how the models performed over time and to see if there were significant changes in performance, we conducted an MRA of the residuals between the modelled and measured flows and determined the significant change points.

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. We note that we did not do this for the hybrid models because part of their construction is based on defining when extreme events occur.

3. Results

3.1 Time-series and model predictive performance

All six models captured well the general pattern and the peaks of the measured flow (Figs 3 and 4). Scatterplots of measured flow against simulated flow, and the associated correlations, are presented in Fig. 5 which, coupled with the calculated indices (Fig. 6), provide a detailed evaluation on the performance of each model. The Modelled H2D and the two Hybrid H2D models produced the largest correlation with measured flow, followed by the Modelled Daily and the Hybrid Daily models. Adding SM as a covariate does not increase model accuracy, as the correlation of Hybrid Daily drops to 0.75 from 0.81 and the correlation of Hybrid H2D drops to 0.84 from 0.91. 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. These results are confirmed by the calculated error and agreement indices. The Modelled H2D and the Hybrid H2D exhibit the smallest error and the greatest agreement with the measured data. Addition of SM as a covariate increases the error and decreases the agreement for both the daily and H2D based models.



319

320 **Fig. 3.** (a) PBM simulated flow at daily resolution (Modelled Daily) and (b) at hourly resolution aggregated to

321 daily (Modelled H2D).

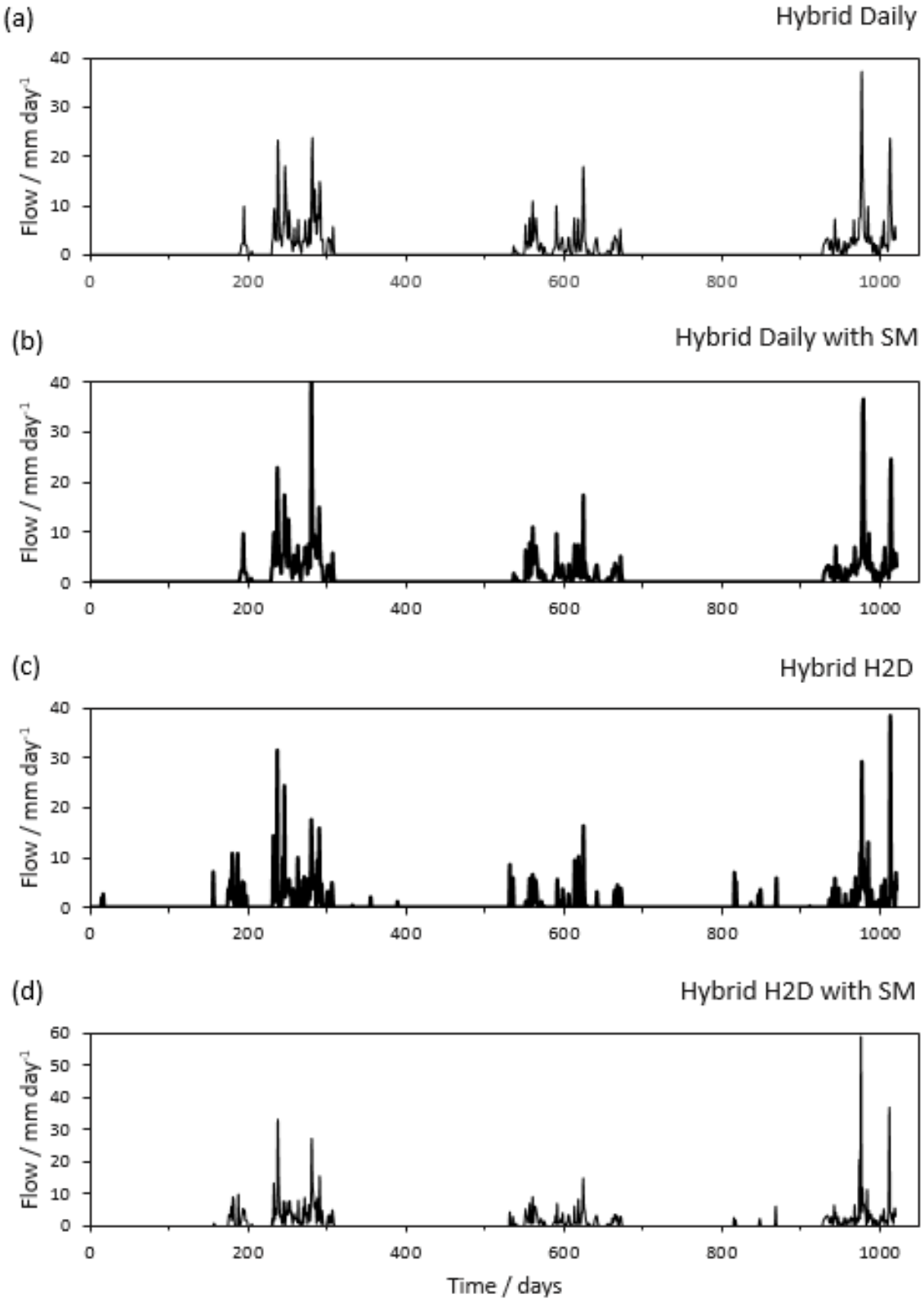


Fig. 4. Hybrid models a) with CEM applied to the maximum daily PBM simulated flow within a peak event and ELM to all other points in the peak event, b) as in (a) but with soil moisture (SM) as a covariate in the ELM model, c) with ELM only applied to the hourly PBM simulated and aggregated to daily flow,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

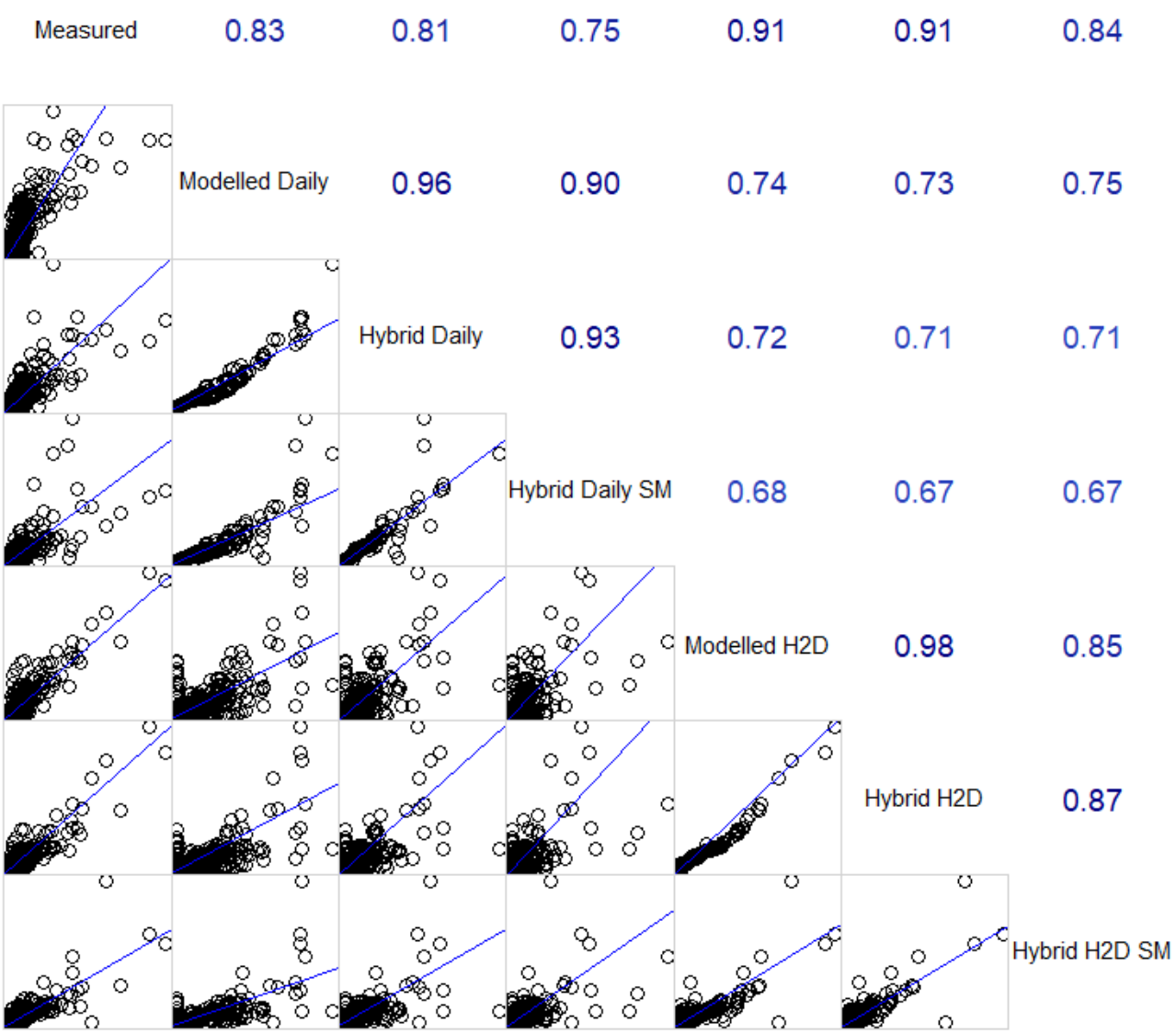


Fig. 5. (Bottom left) scatterplots and (top right) correlations between measured and simulated flow and between flow simulations from the models only.

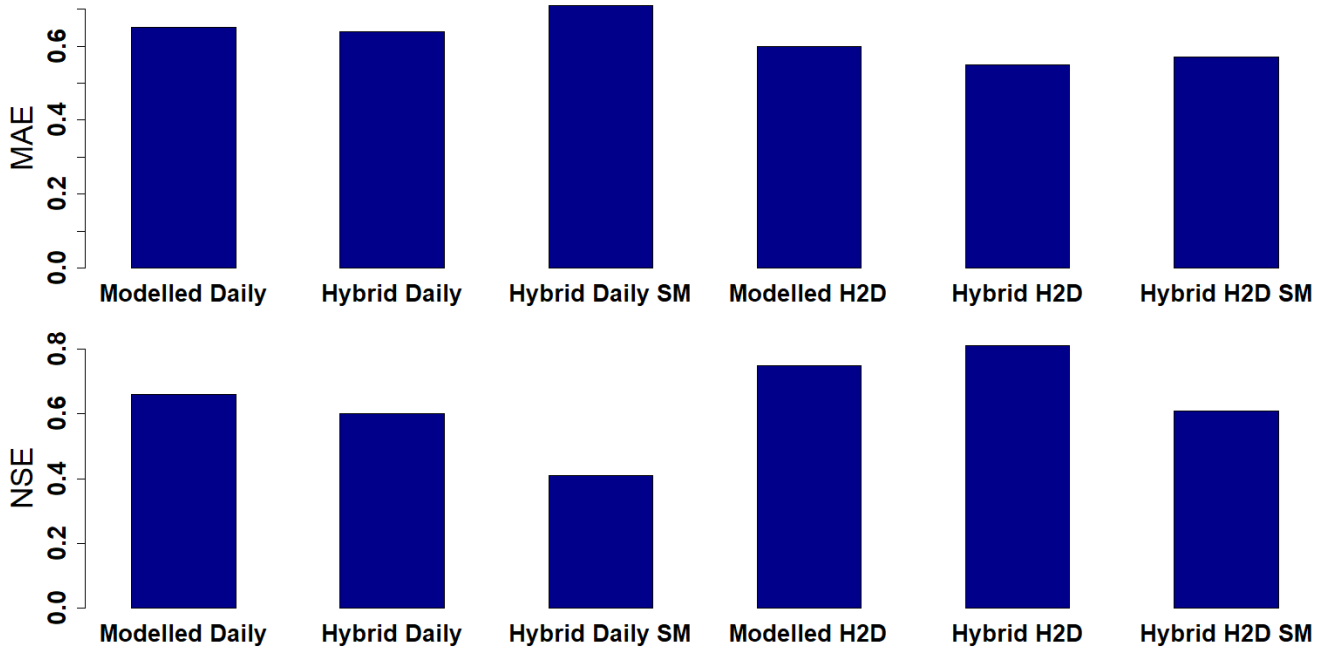


Fig. 6. (top) Mean absolute error (MAE) and (bottom) Nash-Sutcliffe efficiency (NSE) between measured and modelled flow.

3.2 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 (the un-transformed empirical variograms are provided in Appendix A). Authorised variogram models could be fitted to all empirical variograms except for measured flow and SM (Fig. 7). 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 the 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.

Variograms depicted in Fig 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.

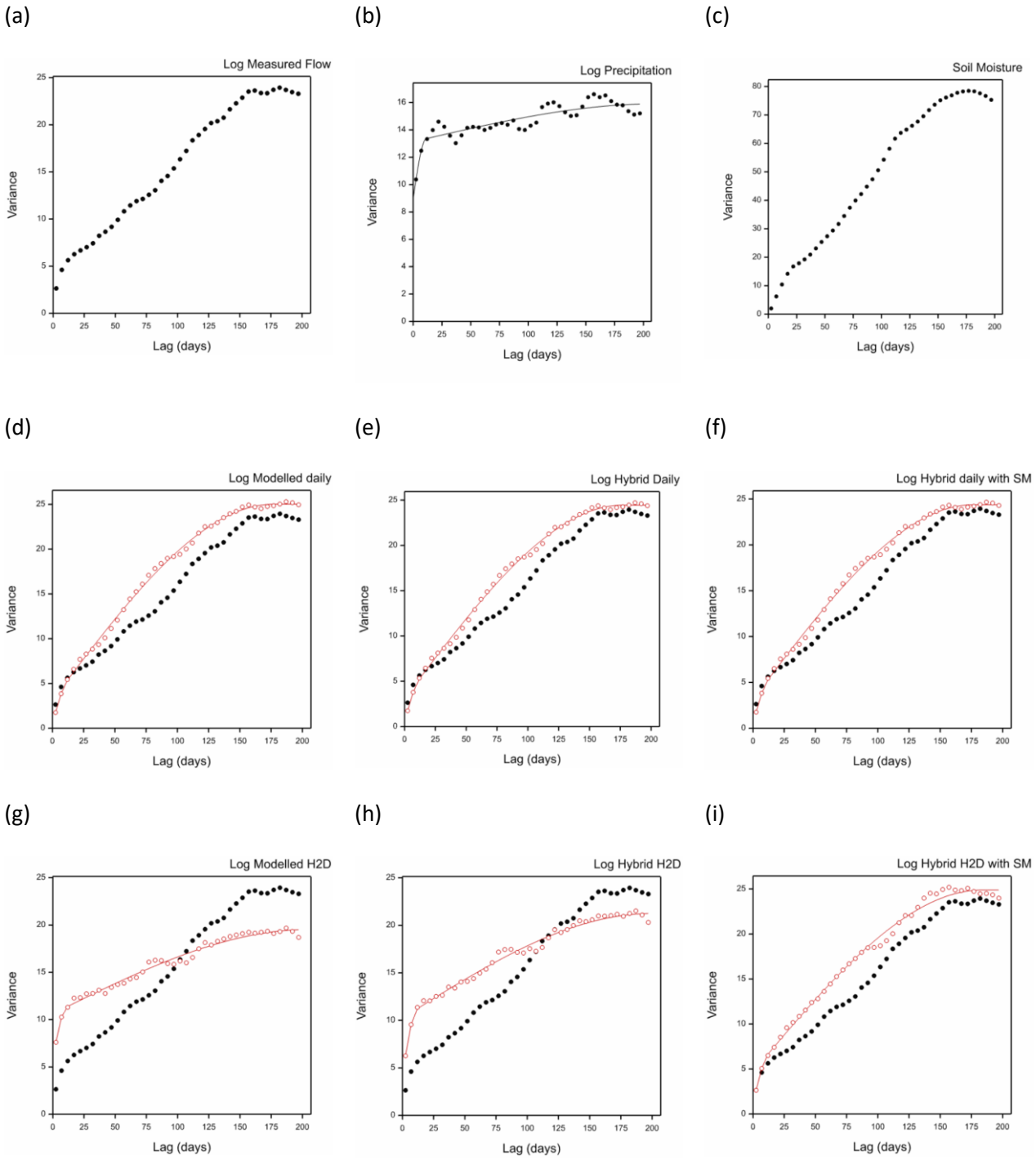
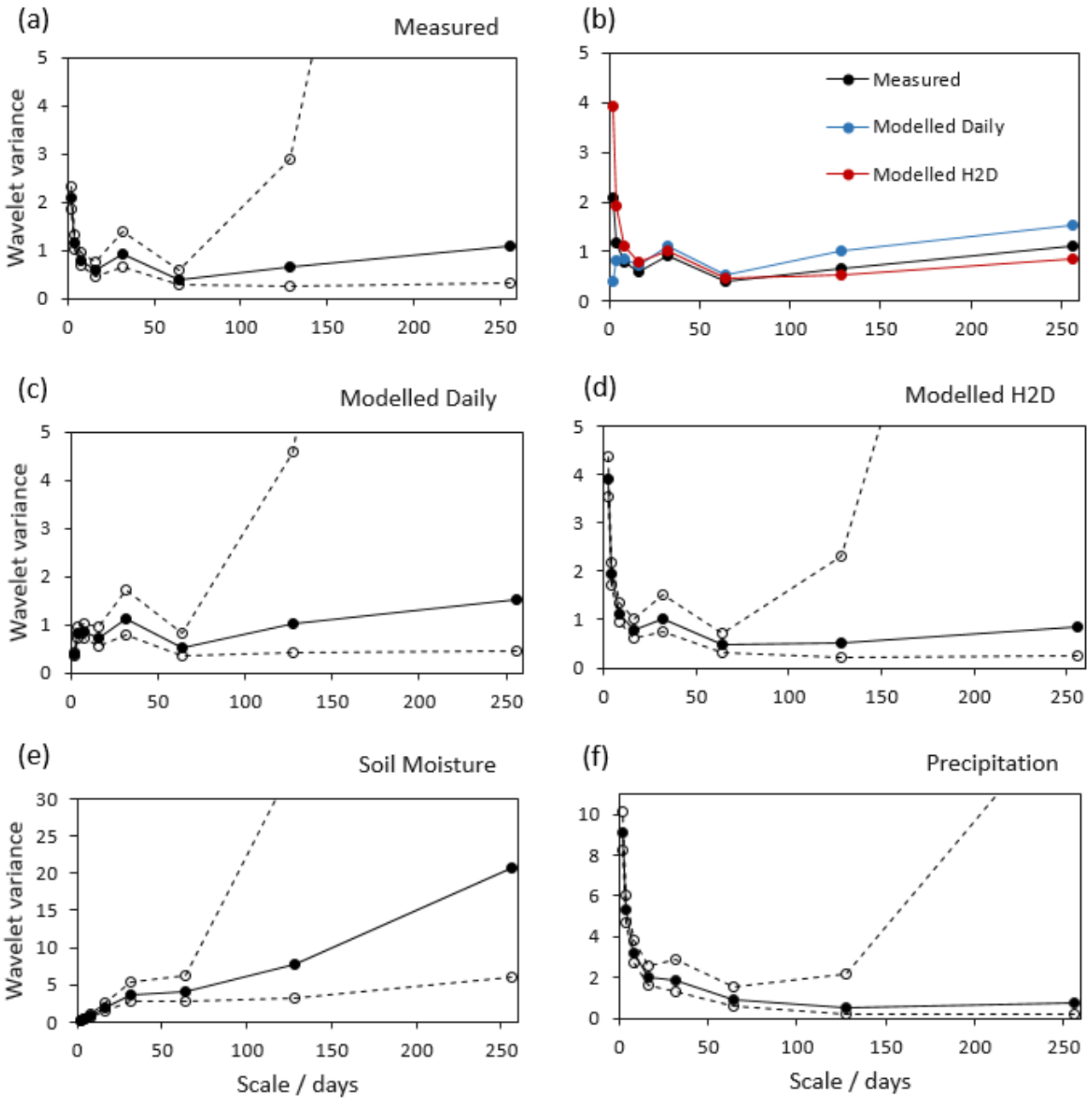


Fig. 7. Empirical variograms of measured (a) log flow, (b) log precipitation and (c) soil moisture. The black line shows the variogram model fitted to the measured data (for precipitation only). Subplots (d-i) show the empirical variograms for log modelled flow variables (red disks) with their respective fitted variogram models (red line) and the empirical variograms of measured log flow for comparison (black disks).

3.3 Wavelet Analysis

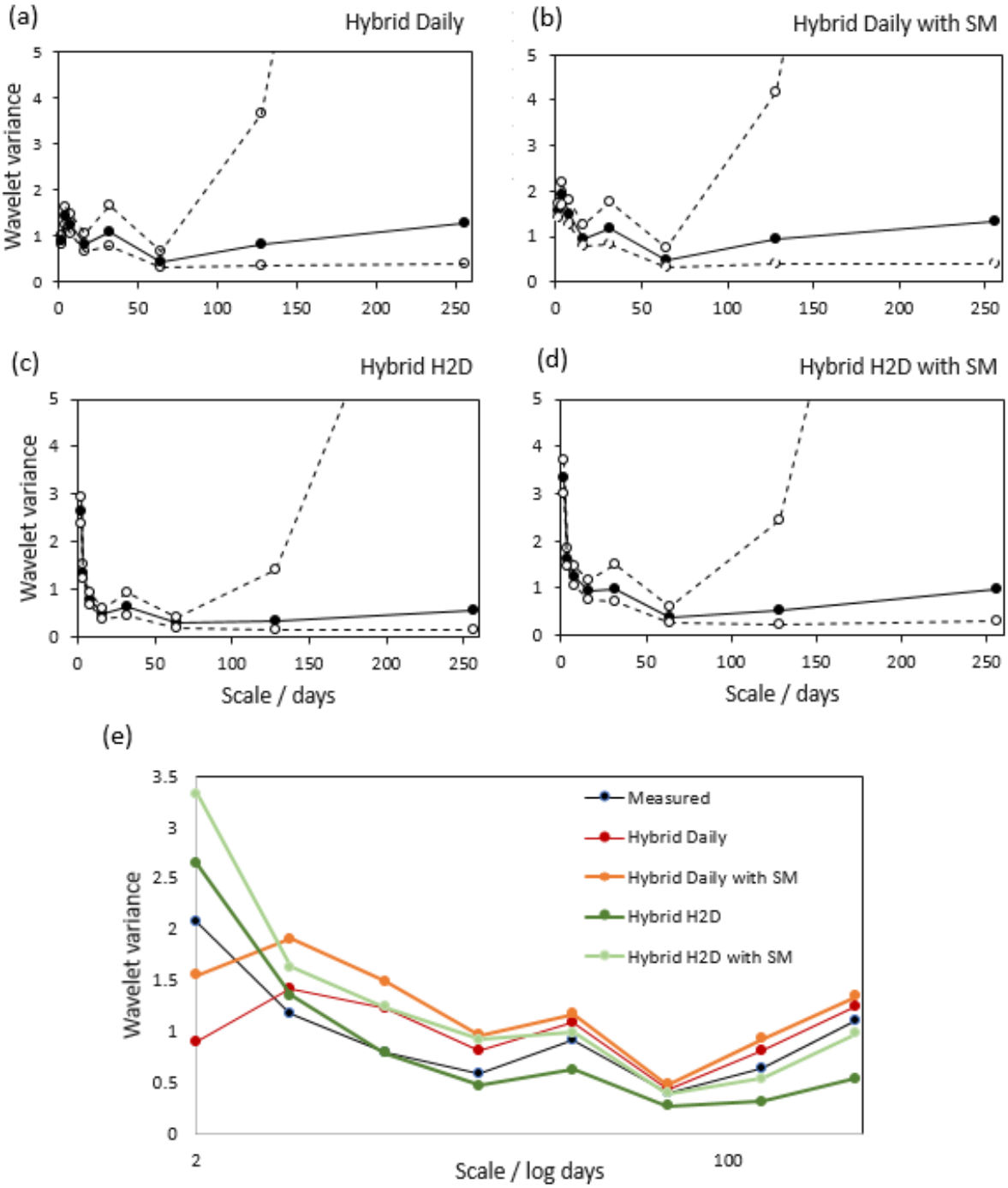
3.3.1 Wavelet variance

The wavelet variance results are given in Figs. 8 and 9. The partition of wavelet variance in the measured 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 (Fig. 8a). Comparing the wavelet variance of the measured data with the PBM simulations (Fig. 8b) 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 (Fig. 9). At coarser scales the Hybrid H2D with SM performs best in this respect (Fig. 9). Using SM as a covariate in the Hybrid Daily model does not increase the accuracy of the predicted variation at coarser scales, however (Fig. 9).



370

371 **Fig. 8.** The wavelet variance for measured (plots a, e, f for flow, SM and precipitation, respectively) and
372 modelled (c and d for Daily and N2D, respectively) data. The wavelet variance is given by the solid discs which
373 mark the lower bound of the scale interval that each wavelet variance is associated with. The open discs
374 show the 95% confidence intervals. The lines are given to aid the eye. Plot (b) compares measured with
375 modelled flow on the same plot.



376

377

378 **Fig. 9.** The wavelet variance for flow simulated with each of the hybrid models. The wavelet variance is given
379 by the solid discs which mark the lower bound of the scale interval that each wavelet variance is associated
380 with. The open discs show the 95% confidence intervals. The lines are given to aid the eye. The bottom plot
381 shows the wavelet variance for all of the hybrid models plotted together with the wavelet variance for the
382 measured data. The scale is presented on the log scale (base 10) to aid inspection of the finer scale variances.

383

3.3.2 MRA of residuals

The MRAs of the residuals for each of the six models are shown in Fig. 10. The significant changes in model performance (as indicated by the red vertical lines) show that the model residuals are greatest 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 1).

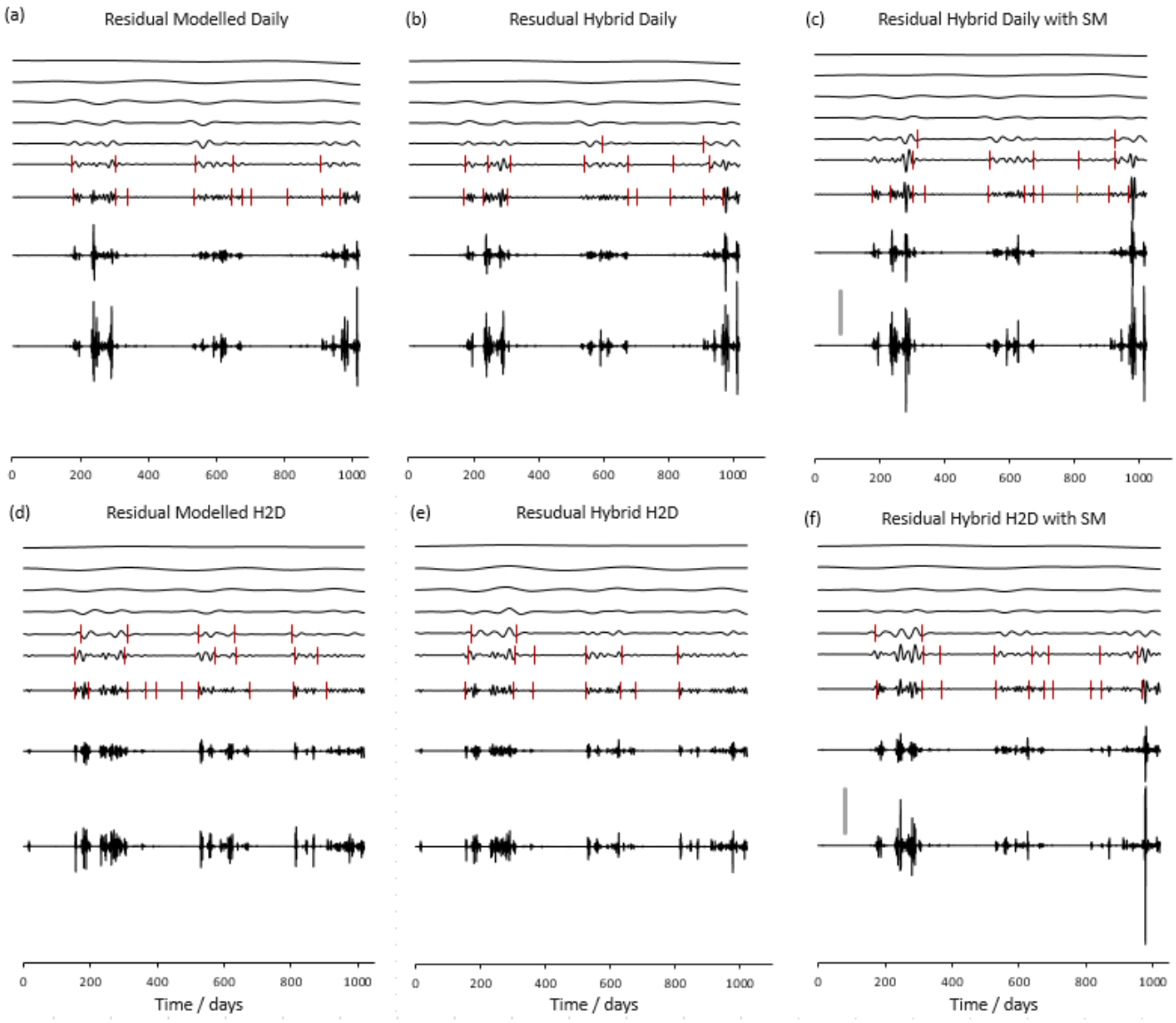


Fig. 10. The MRA for the residuals of each model considered shown as stacked plots. The approximation component is shown at the top of each subplot with variance components plotted below from coarsest at

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 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.

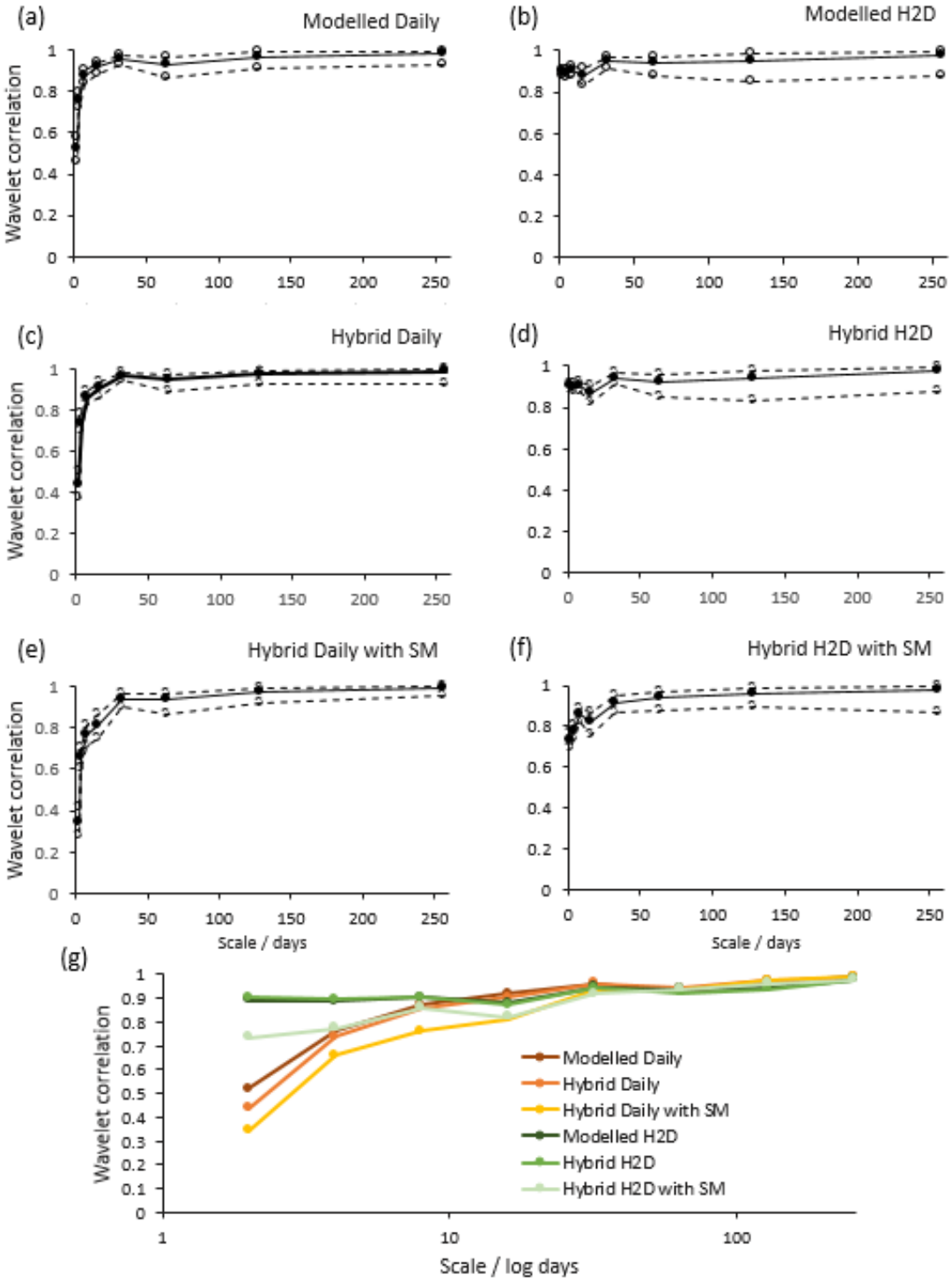
Table 1
The wavelet variances of the residuals for each model.

	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

3.3.3. Wavelet correlations

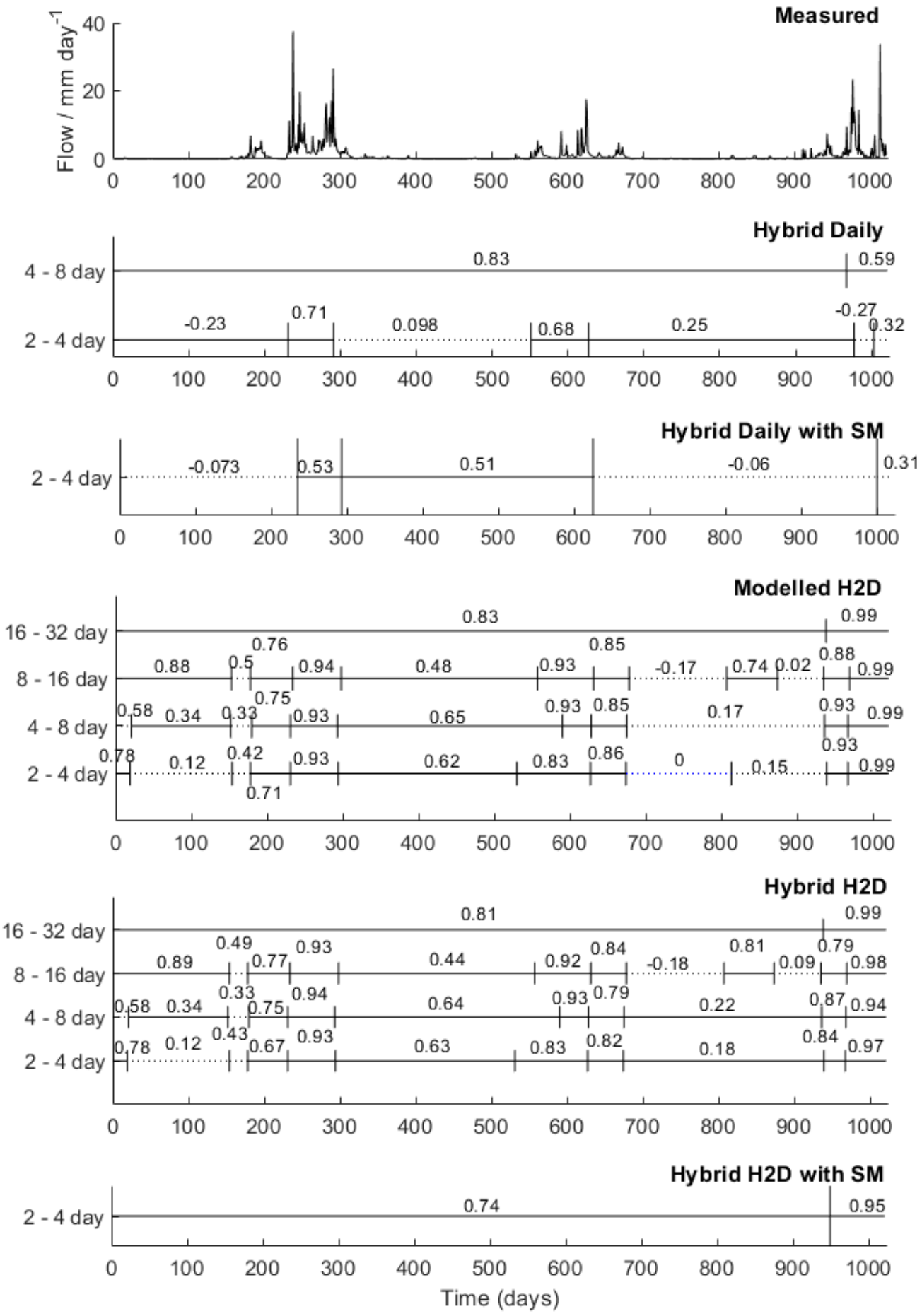
Across the scales, the models derived from PBM flow simulations at the hourly resolution (Modelled H2D, Hybrid H2D and Hybrid H2D with SM) produce a large wavelet correlation with measured flow (>0.7) whereas those based on simulated flow at the daily resolution are less correlated at finer scales (Fig. 11). 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.

412 Significant changes in correlation were detected in the finer scales for the hybrid daily models and
413 the Hybrid H2D with SM, and at scales below 32 days for the Modelled H2D and hybrid H2D (Fig. 12). Broadly
414 speaking, these show that the modelled flow is better correlated with measured during the wet winter
415 periods when high water fluxes are observed. Modelled H2D and Hybrid H2D exhibit a greatest number of
416 changes in the correlation with the measured flow and at the finest scale (2 – 4 day) capture the low flows
417 during dry periods better than the Hybrid Daily models, which show weak correlation with the measured
418 data as during these periods the daily simulations predict no flow. At fine scale, the Hybrid H2D with SM
419 shows a stable correlation of 0.74 for most of the studied period, which increases towards the end of the
420 time series when the SM covariate becomes almost constant.



421

422 **Fig. 11.** The wavelet correlation between simulated and measured flow data. The wavelet correlation is given
423 by the solid discs which mark the lower bound of the scale interval to which each wavelet correlation is
424 associated. The open discs show the 95% confidence intervals. The lines are given to aid the eye. The bottom
425 plot shows the wavelet correlation for all models plotted together. The scale is presented on the log scale
426 (base 10) to aid inspection of the finer scale correlations.


Measured
Flow / mm day-1
Hybrid Daily
4 - 8 day
2 - 4 day
Hybrid Daily with SM
Modelled H2D
16 - 32 day
8 - 16 day
Hybrid H2D
Hybrid H2D with SM
Time (days)

427

Fig. 12. Measured flow (top) and significant changes detected in the wavelet correlation between measured and modelled data at scales where changes were detected. The solid lines indicate that the correlation is significantly different from zero and the dotted that it was not. The Daily Modelled is not depicted as no changes in the wavelet correlation were detected.

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 (Figs. 13a–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 (Figs. 13d–e) (see supplementary information for Modelled H2D).

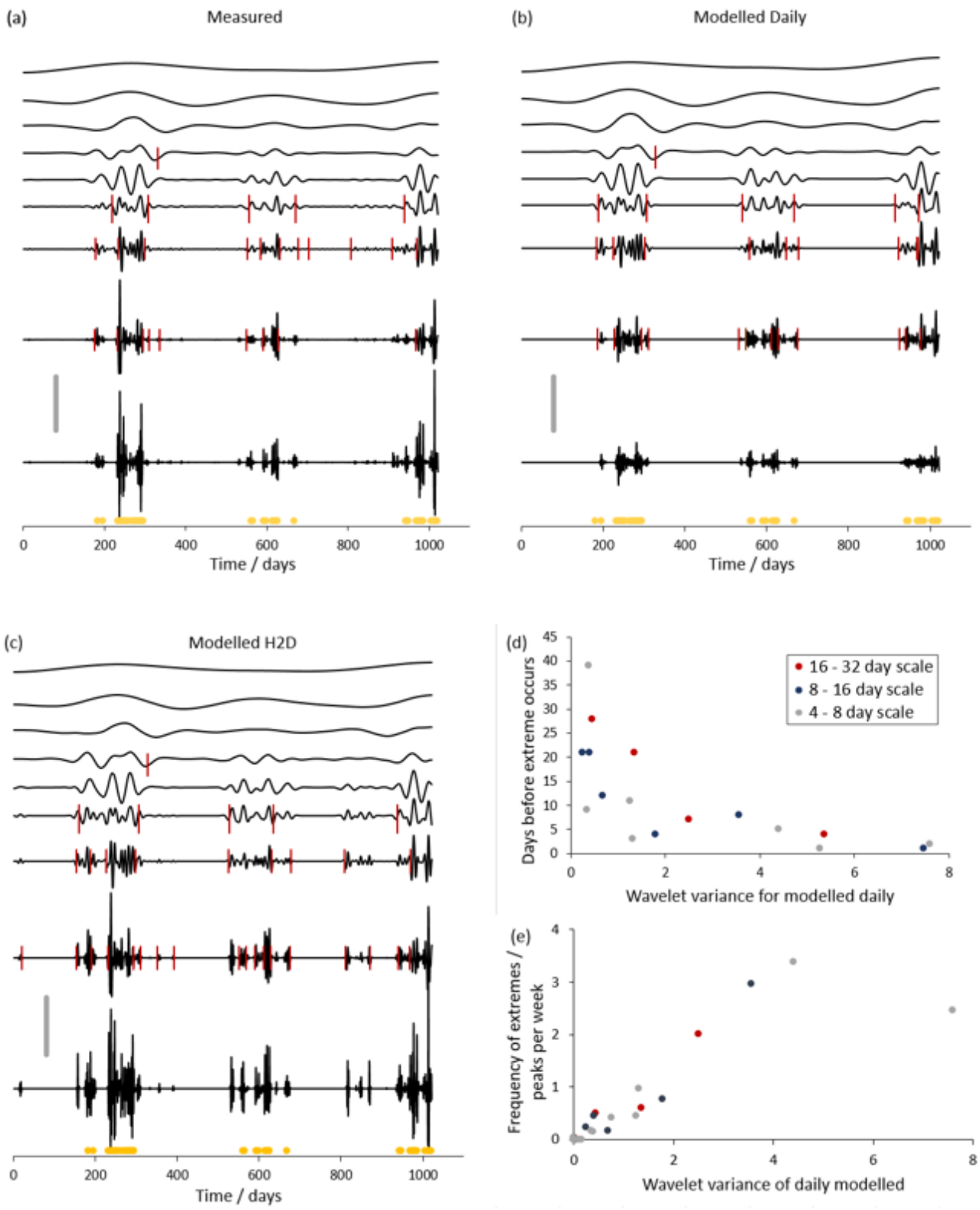


Fig. 13. The MRAs of (a) measured flow, (b) Modelled Daily flow and (c) Modelled H2D flow shown in stacked plots. The approximation component is shown at the top of each subplot with variance components plotted below from coarsest at the top to finest at the bottom. The solid grey scale bar indicates 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.

4. Discussion

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. 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 based on daily resolution simulations (Modelled Daily, Modelled H2D, Hybrid Daily), while the others were based on aggregated hourly resolution simulations (Hybrid Daily with SM, Hybrid H2D, Hybrid H2D with SM). The models were evaluated using a jackknife procedure, where a peak was left out of the training dataset at each iteration. As the whole procedure was based on peak flow events, a split sample could not be applied as an alternative evaluation process.

We explored using diagnostics that are able to reveal how well each model captures the scale dependence in the observed behaviour (wavelet analysis) and how well structural auto-correlation is preserved (variography). 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).

A simple correlation analysis and the calculated indices (Figs 5 and 6) indicated that Modelled H2D and Hybrid H2D were the most accurate predictors of water flow. Both models yielded correlations of r = 0.91, the smallest MAE and the largest NSE. Surprisingly, the inclusion of SM provided no additional predictive information. The wavelet analysis reveals more information about this showing that the SM covariate negatively impacts the correlation at finer scales in particular. The wavelet correlation change detection

475 reinforces the fact that the fine scale predictions suffer from inclusion of the SM covariate showing the largest
476 correlation between Hybrid H2D with SM and measured when the SM covariate is almost flat and so offering
477 negligible predictive power (Figs 2 and 12). The fact that SM has no positive effect on the models'
478 performance could have several explanations. Measuring SM is known to be more difficult compared, for
479 example, to measuring precipitation. Therefore, a greater uncertainty in the SM measurements is likely (see
480 below). Moreover, the flow is representative over the whole sub-catchment gathered at the flume whereas
481 SM (and precipitation) is measured at only one point and so may not be as representative of catchment-scale
482 SM. The relatively poor model performance may also result from overfitting to the training dataset.

483 It is clear from the scatterplots of Fig. 5 that there are issues of under-prediction of peak flows
484 associated with models derived from the Daily PBM simulation. This is reflected in the wavelet variance
485 where it is evident that the fine-scale wavelet variance is underestimated (Fig. 8b). The hybrid approach
486 mitigates this effect to some extent, but variation is still smaller than it should be at the fine scale (bottom
487 plot in Fig. 9). In all three Daily-based models, the relatively small fine-scale wavelet variation is
488 overcompensated for at mid-to-coarse scales. Conversely, the H2D-based models tend to overestimate fine-
489 scale variation (Figs. 8b and 9) with the most extreme effects seen in Modelled H2D (Fig. 8b). The hybrid
490 models dampen this overestimation in the H2D-based models with Hybrid H2D capturing the fine-scale
491 variation the best out of all six models. Hybrid H2D also shows the overall best wavelet correlation at finer
492 scales (<32 days), while at coarser scales (32> days) all models produce a large wavelet correlation with the
493 measured data (Fig. 11). Thus, for Hybrid H2D, this complements the high performance of its standard
494 correlation with the measured data.

495 The variography offers more evidence about the effect of over predicting the fine-scale variation.
496 First we note that for variography we chose to log transform the measured and modelled flow data.
497 Asymmetry or skewness in data generally has little effect on variogram estimation for large samples, and so
498 predictions may usually safely be done with the raw data (Webster and Oliver, 2007). However, in our case
499 we found that a "hole effect" in the empirical variogram meant it was not possible to fit a valid variogram
500 model (known as an authorised model in geostatistical literature) without transformation. Transformation
501 does, however, dampen the extremes in the data. Therefore, for variography we compared only the

variogram models between measured and modelled variants. Comparing the variograms of the modelled and measured flow data (Fig. 7) 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. 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. This is likely to be due to the smoothing effect of this covariate, which notably has its largest component of variation at course scale (Fig. 8)

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 185 days (see S1 for the variogram model parameters). In each case, the double spherical model was found to be the best fitting model supporting further that there are two substantial sources of variation in the data. All modelled variograms could capture the short- (≈12-days) and the long-range processes (≈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. For SM, the short- and long-range components were approximately 20 and 175 days.

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 (Fig. 10). Rust et al., showed that change detection methods were able to identify significant effects of land use change. We have no similar effect here as the catchment was managed in a consistent way across the timeseries, however we do see significant changes at finer scales that relate to periods of increased flow. It is evident from the residuals and wavelet correlation that the model performance is not consistent across time and that, in particular, the Daily-based models perform less well over the last major burst of activity (900 days onward). This corresponds with a period where the soil is quite saturated according to the

529 measured data and so suggests that this local measurement of soil moisture and the daily modelled
530 predictions do not capture the more complex soil-water dynamics that operate across the sub-catchment
531 and in this case, dampen flow.

532 The extreme events identified using the automated threshold stability method (as given in Curceac
533 et al., 2020a; 2020b) did somewhat accord with the wavelet change point detection analysis (Fig. 13). The
534 local wavelet variance of the model predictions (i.e., only those from the Modelled Daily and Modelled H2D)
535 was correlated with the number of extreme events and a large wavelet variance suggested that extreme
536 events were imminent. The wavelet-based method was less efficient for predicting extremes, than simply
537 applying the automated threshold stability method to the model prediction. However, it serves well in an
538 exploratory and complementary context.

539 ## 5. Conclusions

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

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

555 complex and so evaluating the risk of extreme water flows and associated flooding will continue to require
556 complex solutions.

557 **Acknowledgements**

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

563 The study datasets are freely available from https://www.rothamsted.ac.uk/north-wyke-farm-platform and
564 the SPACSYS (PBM) model can be found here https://www.rothamsted.ac.uk/rothamsted-spacsys-model.
565 The variogram analysis was conducted in GENSTAT (VSN International, 2019), while R software (R Core Team,
566 2019) was used for the implementation of the hybrid models, where the CEM is from the texmex R package
567 (Southworth et al., 2020) and the ELM from the elmNNRcpp R package (Mouselimis and Gosso, 2020).

568

569 **References**

570 Bates, B. C., Kundzewicz, Z. W., Wu, S. and Palutikof, J. P. (2008). Climate Change and Water. Technical
571 Paper of the Intergovernmental Panel on Climate Change, IPCC Secretariat, Geneva, 210 pp.
572 Bogner, Konrad, Katharina Liechti, and Massimiliano Zappa. 2016. "Post-Processing of Stream Flows in
573 Switzerland with an Emphasis on Low Flows and Floods." *Water* 8 (4): 115.
574 https://doi.org/10.3390/w8040115.
575 Bouraoui, F., Grizzetti, B., Granlund, K., Rekolainen, S. and Bidoglio, G. (2004). Impact of Climate Change on
576 the Water Cycle and Nutrient Losses in a Finnish Catchment, Climatic Change, 66(1-2), 109-126. doi:
577 10.1023/B:CLIM.0000043147.09365.e3.
578 Brown, I., Bardgett, R., Berry, P., Crute, I., Morison, J., Morecroft, M., Pinnegar, J., Reeder, T. and Topp. K.
579 (2016). UK Climate Change Risk Assessment, Chapter 3: Natural Environment and Natural Assets.
580 Chilès, J. P. and Delfiner P. (2009). Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons.
581 Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer, London, UK.
582 Comber, A., Brunsdon, C., Charlton, M. and Harris, P. (2017). Geographically Weighted Correspondence
583 Matrices for Local Error Reporting and Change Analyses: Mapping the Spatial Distribution of Errors
584 and Change, Remote Sensing Letters, 8(3), 234-243. doi: 10.1080/2150704X.2016.1258126.
585 Curceac, S., Atkinson, P. M., Milne, A, Wu, L. and Harris, P. (2020a). Adjusting for Conditional Bias in Process
586 Model Simulations of Hydrological Extremes: An Experiment Using the North Wyke Farm Platform,
587 Frontiers in Artificial Intelligence, 3. doi: 10.3389/frai.2020.565859.
588 Curceac, S., Atkinson, P. M., Milne, A., Wu, L. and Harris, P. (2020b). An Evaluation of Automated GPD
589 Threshold Selection Methods for Hydrological Extremes across Different Scales, Journal of
590 Hydrology, 585, 124845. doi: 10.1016/j.jhydrol.2020.124845.
591 Daubechies, I. (1988). Orthonormal Bases of Compactly Supported Wavelets, Communications on Pure and
592 Applied Mathematics, 41(7), 909-996. doi: 10.1002/cpa.3160410705.
593 Deo, R. C. and Şahin, M. (2016). An Extreme Learning Machine Model for the Simulation of Monthly Mean
594 Streamflow Water Level in Eastern Queensland, Environmental Monitoring and Assessment, 188,
595 90. doi: 10.1007/s10661-016-5094-9.
596 Field, C. B., Barros, V., Stocker, T. F. and Dahe, Q. (2012). Managing the Risks of Extreme Events and
597 Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on
598 Climate Change, Cambridge, Cambridge University Press. doi: 10.1017/CBO9781139177245.
599 Goovaerts, P. 1997. Geostatistics for Natural Resource Evaluation, Technometrics, Vol. 42.
600 Gringarten, E. and Deutsch, C. V. (2001). Teacher's Aide Variogram Interpretation and Modeling,
601 Mathematical Geology, 33(4), 507-534. doi: 10.1023/A:1011093014141.
602 Harris, P., Brunsdon, C. and Charlton, M. (2013). The Comap as a Diagnostic Tool for Non-Stationary Kriging
603 Models, International Journal of Geographical Information Science, 27(3), 511-541. doi:
604 10.1080/13658816.2012.698014.
605 Heffernan, J. E. and Tawn, J. A. (2004). A Conditional Approach for Multivariate Extreme Values (with
606 Discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(3), 497-
607 546. doi.org: 10.1111/j.1467-9868.2004.02050.x.
608 Huang, G. B., Zhu, Q. Y. and Siew, C. K. (2006). Extreme Learning Machine: Theory and Applications,
609 Neurocomputing, Neural Networks, 70(1), 489-501. doi: 10.1016/j.neucom.2005.12.126.
610 Jaiswal, R. K., Ali, S. and Bharti, B. (2020). Comparative Evaluation of Conceptual and Physical Rainfall–
611 Runoff Models, Applied Water Science, 10(1), 48. doi: 10.1007/s13201-019-1122-6.
612 Keef, C., Papastathopoulos, I. and Tawn, J. A. (2013). Estimation of the Conditional Distribution of a
613 Multivariate Variable given That One of Its Components Is Large: Additional Constraints for the
614 Heffernan and Tawn Model, Journal of Multivariate Analysis, 115, 396-404. doi:
615 10.1016/j.jmva.2012.10.012.
616 Kundzewicz, Z. W., Mata, L. J., Arnell, N. W., Doll, P., Kabat, P., Jimenez, B. et al. (2007). Freshwater
617 Resources and Their Management. In Climate Change 2007: Impacts, Adaptation and Vulnerability.
618 Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel
619 on Climate Change, edited by M. L. Parry, O. F. Canziani, J. P. Palutikof, P. J. van der Linden, and C.
620 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.

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

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.

Papacharalampous, Georgia, Hristos Tyralis, Andreas Langousis, Amithirigala W. Jayawardena, Bellie Sivakumar, Nikos Mamassis, Alberto Montanari, and Demetris 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.

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.

675 Webster, R., and Oliver, M. A. (2007). *Geostatistics for Environmental Scientists,* John Wiley & Sons.
676 Wu, L., McGechan, M. B. McRoberts, N., Baddeley, J. A. and Watson, C. A. (2007). SPACSYS: Integration of a
677 3D Root Architecture Component to Carbon, Nitrogen and Water Cycling-Model Description,
678 Ecological Modelling, 200(3), 343-359. doi: 10.1016/j.ecolmodel.2006.08.010.
679 Yaseen, Z. M., Sulaiman, S. O., Deo, R. C. and Chau, K. W. (2019). An Enhanced Extreme Learning Machine
680 Model for River Flow Forecasting: State-of-the-Art, Practical Applications in Water Resource
681 Engineering Area and Future Research Direction, Journal of Hydrology, 569, 387-408. doi:
682 10.1016/j.jhydrol.2018.11.069.
683

684 **Supplementary 1**

685 The base models considered for variogram models were

686 Spherical:

687
$$\gamma(h) = c_0 + c\left\{\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right\} \text{ for } h \le a$$
$$= c_0 + c \text{ for } h > a$$
$$= 0 \text{ for } h = 0,$$

688 where h is a scalar in temporal distance only. Its parameters are c_0 which is the nugget variance, c is the
689 correlated variance and a is the distance parameter (the range) of the model. Parameter a is the limiting
690 distance of temporal dependence or correlation. The parameter c is the variance of the correlated structure,
691 so that $c_0 + c$ is the total variance of the underlying random process, of which the data are a realization.

692

693 Circular:

694
$$\gamma(h) = c_0 + c\left\{1 - \frac{2}{\pi}\cos^{-1}\left(\frac{h}{a}\right) + \frac{2h}{\pi a}\sqrt{1 - \frac{h^2}{a^2}}\right\} \text{ for } h \le a$$
$$= c_0 + c \text{ for } h > a$$
$$= 0 \text{ for } h = 0,$$

695 in which the parameters c_0, c and a are defined in the same way as for the spherical model.

696

697 Exponential model:

698
$$\gamma(h) = c_0 + c\left\{1 - \exp\left(-\frac{h}{r}\right)\right\}$$

699 in which the parameters c_0 and c are defined as above but r is the distance parameter which is approximately

700 a third of the effective range (see Webster and Oliver, 2007).

701 For each of our variables the double spherical proved the best model. This is given by:

$$\gamma(h) = c_0 + c_1\left\{\frac{3h}{2a_1} - \frac{1}{2}\left(\frac{h}{a_1}\right)^3\right\} + c_2\left\{\frac{3h}{2a_2} - \frac{1}{2}\left(\frac{h}{a_2}\right)^3\right\} \text{ for } h \le a_1$$

702
$$= c_0 + c_1 + c_2\left\{\frac{3h}{2a_2} - \frac{1}{2}\left(\frac{h}{a_2}\right)^3\right\} \text{ for } a_1 < h \le a_2$$
$$= c_0 + c_1 + c_2 \text{ for } h > a_2$$

703
$$= 0 \text{ for } h = 0,$$

704 The parameters for each model are given in the table below:

	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

705

706

Appendix A

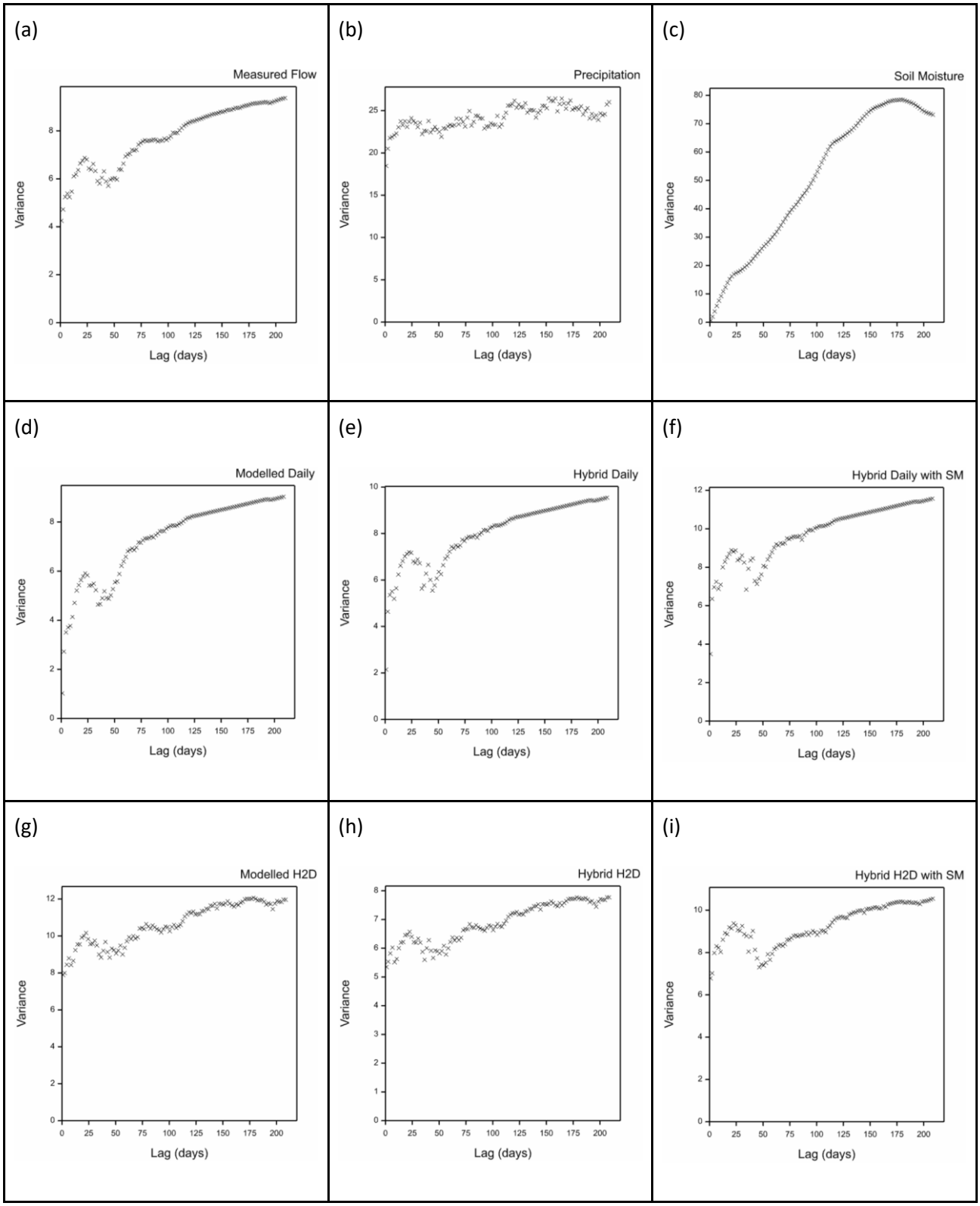


Fig. A.1. Empirical variograms of measured (a) flow, (b) precipitation, (c) soil moisture and (d-i) modelled flow variable.