Front. Earth Sci. 2017, 11(2): 203–213
DOI 10.1007/s11707-017-0622-7

RESEARCH ARTICLE

Simple statistical models for relating river discharge with precipitation and air temperature—Case study of River Vouga (Portugal)

T. STOICHEV (✉)[1], J. ESPINHA MARQUES[2], C.M. ALMEIDA[1], A. DE DIEGO[3], M.C.P. BASTO[4], R. MOURA[2], V.M. VASCONCELOS[1]

1 Interdiscplinary Center of Marine and Environmental Research (CIIMAR/CIMAR), University of Porto, 4450-208 Matosinhos, Portugal
2 Institute of Earth Sciences (ICT) and Department of Geosciences, Environment and Land Planning, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal
3 Department of Analytical Chemistry, Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48080 Bilbao, Basque Country, Spain
4 CIIMAR/CIMAR and Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal

Abstract Simple statistical models were developed to relate available meteorological data with daily river discharge (RD) for rivers not influenced by melting of ice and snow. In a case study of the Vouga River (Portugal), the RD could be determined by a linear combination of the recent (P_R) and non-recent (P_{NR}) atmospheric precipitation history. It was found that a simple linear model including only P_R and P_{NR} cannot account for low RD. The model was improved by including non-linear terms of precipitation that accounted for the water loss. Additional improvement of the models was possible by including average monthly air temperature (T). The best model was robust when up to 60% of the original data were randomly removed. The advantage is the simplicity of the models, which take into account only P_R, P_{NR} and T. These models can provide a useful tool for RD estimation from current meteorological data.

Keywords multiple regression, atmospheric precipitation, river discharge, runoff, Aveiro Lagoon

1 Introduction

Coastal lagoons are very complex ecosystems under the influence of oceanic and continental environments: inflow and outflow of saltwater (from the oceans) and freshwater (from rivers and aquifers), multiple physical and chemical influences from the adjacent terrestrial and aquatic ecosystems and several anthropological impacts. In order to better understand the chemical and biological processes in lagoons and estuaries it is important to know the hydrology of the system (Dias et al., 1999).

However, daily river discharge (RD) recording may be interrupted and, consequently, recent data on RD are often not available, thereby posing great challenges for hydrological studies. Several types of models, which may be classified as lumped or distributed and also as deterministic or stochastic (Beven, 2012) have been applied for estimating and predicting RD (Hurkmans et al., 2008; Achleitner et al., 2012).

The accurate simulation of RD demands a thorough characterization of the most significant hydrological processes acting in the catchment and therefore requires a substantial amount of data regarding topography, geology, hydropedology, land cover, and climate (Xu et al., 2011; Vilaysane et al., 2015). Also, a solid conceptual model is necessary in order to obtain realistic results. A more recent approach relies on models based on fuzzy logic, which seem to be better able to represent the high complexity of hydrological systems (Sen, 2010).

Nevertheless, hydrological research must often be carried out in a context of data shortage. In fact, detailed information regarding the catchment features may not be available for a number of reasons. For example, the application of fully physically-based models may require extensive field surveys to quantify soil, vegetation, and geological features, among other aspects. Such surveys are frequently only feasible in small and massively instrumented experimental catchments (Beven, 1989).

Received January 27, 2016; accepted October 9, 2016

E-mail: tstoichevbg@yahoo.com

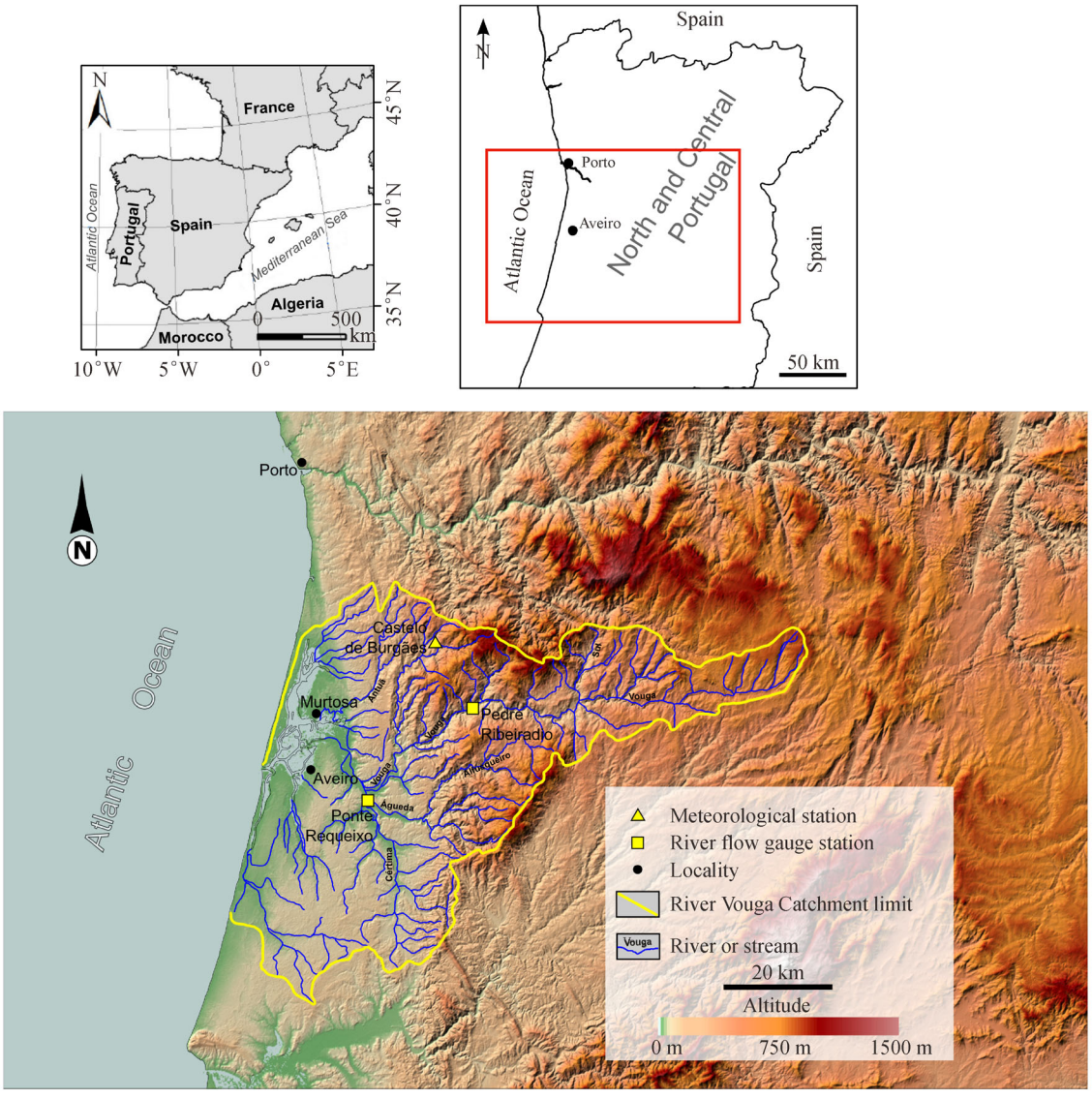

Fig. 1 Regional framework of the Aveiro Lagoon. DEM obtained from the Shuttle Radar Topography Mission (SRTM), in 90 meter resolution, produced by NASA and distributed by the United States Geological Survey [USGS], 2009.

This circumstance has motivated the development of simple models that demand less input but still are able to provide satisfactory results (Muthukrishnan et al., 2006; Du et al., 2007). In spite of the existence of more complex physical models, RD may be estimated from the historical input (precipitation, air temperature) and output (RD) which are often more easily available (Brath et al., 2004). However, on a daily time scale, the RD may present large variations and its estimation could be a challenging task. In order to overcome this difficulty one may use longer periods, for example, monthly RD. This aspect may be critical in small mountain basins (which tend to have lower concentration times) especially if the discharge measurements are made only once a day (Espinha Marques et al., 2011).

This study is an attempt to use only readily available meteorological data to estimate the daily RD. The Aveiro Lagoon (so called "Ria de Aveiro") was considered as a case study. It is a coastal, biologically productive lagoon, located in Central Portugal (Fig. 1) (Cerejo and Dias, 2007) receiving freshwater mainly from the Vouga River (Dias et al., 2007). Therefore, the RD of Vouga River downstream was related to the average daily precipitation (P) and average monthly temperatures (T) available during the same period. Several simple models were developed to study the influence of the precipitation history, with and without consideration of T. The developed models in this study do not account for the complex phenomena that might occur in the drainage area of the Vouga River. Their strongest point is their simplicity (using only meteorolo-

gical data for P and T) and high robustness. Due to a relatively small drainage area and the simultaneous availability of hydrological and meteorological data, only one meteorological station was used. For rivers with larger drainage basins, precipitation data from several meteorological stations should be considered.

2 Study region

The Vouga River catchment is situated in Central Portugal (Fig. 1) and covers an area of around 3362 km^2. The Vouga River source is located at 930 m a.s.l. in the Lapa Mountain and has a total length of around 141 km. The RD to the Aveiro Lagoon is mainly from the Vouga River (50 $m^3 \cdot s^{-1}$ average flow) which represents 2/3 of the freshwater entering the lagoon (Dias et al., 2007). The Antuã River is the second largest source whose average flow is only 5 $m^3 \cdot s^{-1}$. There are other rivers entering the Aveiro Lagoon (Sul, Caima, Cértima and Alfusqueiro tributaries) but their importance is negligible.

The Eastern part of the basin comprises metamorphic and igneous terrains consisting mostly of Pre-Ordovician metasedimentary rocks and Variscan granitic rocks (Oliveira et al., 1992). The western part of the basin consists mostly of Mesocenozoic sedimentary rocks.

The geologic nature of the catchment encompasses a great geomorphological contrast. The eastern sector of the catchment has an irregular relief, often of mountainous nature with altitude reaching around 1000 m a.s.l., while the western sector, which borders the lagoon, has smoother relief and altitude much closer to sea level.

According to Agencia Estatal de Meteorologia (Spain) and Instituto de Meteorologia (Portugal) (AEMET-IM, 2011) the Vouga River catchment Köppen-Geiger climate classification is Csb which corresponds to a temperate with dry or temperate summer climate. The spatial distribution of precipitation and air temperature values vary significantly across the catchment due to its location regarding the Atlantic ocean and the contrasting relief and altitude. Mean annual precipitation ranges from less than 1000 mm/yr, in the lagoon surrounding plains, to more than 1500 mm/yr, in the higher mountain areas (Van der Weijden and Pacheco, 2006). The T at Aveiro (1981 to 2010 data) ranges from a minimum average value of 10.1°C (in January) to a maximum average value of 20.4°C (in August) (Instituto Português do Mar e da Atmosfera [IPMA], 2015).

In the eastern sector of the basin the soil types are Cambisol, Umbrisol, Regosol, Anthrosol, and Leptosol, while in the western sector the soil types are Solonchak, Fluvisol, Umbrisol, Antrosol, and Cambisol (Rogado et al., 1992; Agroconsultores and Geometral, 1995; European Soil Bureau Network [ESBN], 2005).

Agriculture (small-size farmlands) and forest are the prevailing land use in the catchment (around 4/5 of the total area). The remaining part consists of rock outcrops (in the mountains), urban areas, and water bodies (Van der Weijden and Pacheco, 2006).

3 Methods

3.1 Data collection

All historical data were taken from the Sistema Nacional de Informação de Recursos Hídricos, Portugal (http://snirh.pt/). The RD was determined as the sum from Pedre Ribeiradio (09H/01H) discharge (Vouga River) and Ponte Requeixo (10F/02H) discharge (Agueda River) for the period 10/03/1978–30/09/1980 (Fig. 1). Meteorological data for P (mm) and T (°C) were taken from Barragem de Castelo Burgães station (08G/01C, Fig. 1) for the period 04/02/1978–30/09/1980.

3.2 Statistical methods

The historical data were treated by means of a linear model using R (R Core Team, 2014). The dependent variable (RD) was normalized using the Box-Cox function (Crawley, 2007; Bellanger and Tomassone, 2014):

$$RD_T = \frac{RD^{\lambda} - 1}{\lambda},$$

where RD_T is the transformed dependent variable and λ is the parameter of the transformation.

In all models the precipitation history is separated into recent and non-recent events. The recent precipitation events (P_R) represent daily precipitations P_i, averaged for periods of one, three, or five days before the day i for which the RD_T should be calculated. These explanatory variables are denoted as $P1$, $P3$, and $P5$, respectively. The non-recent precipitation events (P_{NR}) are daily precipitations P, averaged for ten or thirty day periods prior to each of the studied recent precipitation periods. They were denoted $P2\_11$ and $P2\_31$ (preceding recent precipitation period for $P1$), $P4\_13$, and $P4\_33$ (preceding recent precipitation period for $P3$), $P6\_15$ and $P6\_35$ (preceding recent precipitation period for $P5$). Therefore, six fits were developed within each model using P_R averaged for 1, 3, or 5 day periods and P_{NR} averaged for 10 or 30 day periods, respectively.

In linear model 1, RD_T was calculated as a linear combination of only P_R and P_{NR} without T:

$$RD_T = a_0 + a_1 P_R + a_2 P_{NR} + a_{1,2} P_R P_{NR}. \quad (1)$$

In linear model 2, as in model 1, the RD_T was calculated using only data from precipitation history but quadratic terms of both of P_R and P_{NR} were also included:

$$RD_T = a_0 + a_1 P_R + a_2 P_{NR} + a_{1,2} P_R P_{NR} + a_{1,1} P_R^2 + a_{2,2} P_{NR}^2. \quad (2)$$

Linear model 3 includes a combination of P_R, P_{NR} and T:

$$RD_T = a_0 + a_1 P_R + a_2 P_{NR} + a_3 T + a_{1,2} P_R P_{NR} + a_{1,3} P_R T + a_{2,3} P_{NR} T + a_{1,2,3} P_R P_{NR} T. \quad (3)$$

Linear model 4 includes a combination of P_R, $P_{NR,}$ and T but has quadratic terms of P_R and P_{NR} (as in model 2):

$$RD_T = a_0 + a_1 P_R + a_2 P_{NR} + a_3 T + a_{1,2} P_R P_{NR} + a_{1,3} P_R T + a_{2,3} P_{NR} T + a_{1,2,3} P_R P_{NR} T + a_{1,1} P_R^2 + a_{2,2} P_{NR}^2. \quad (4)$$

In the above equations, the coefficients a_0 represent the intercept, a_1, a_2, a_3 – the simple effects of P_R, $P_{NR,}$ and T and $a_{1,1}$, $a_{2,2}$ – the quadratic effects of the P_R and P_{NR}, respectively. The coefficients $a_{1,2}$, $a_{1,3}$, $a_{2,3}$ account for the double interactions $P_R \times P_{NR}$, $P_R \times T$, and $P_{NR} \times T$, respectively and $a_{1,2,3}$ – for the triple interaction $P_R \times P_{NR} \times T$. These coefficients were determined using multiple regression analysis (Crawley, 2007). The models were simplified by leaving only coefficients significantly different from 0 ($p < 0.05$) by gradual removal of non-significant coefficients, starting with the most complex interaction terms.

The obtained best adequate models were compared on the basis of their success and failure characteristics. The frequency of success of a model was determined as the fraction of times (%) when the calculated RD_{MODEL} differed no more than twice (higher or lower) from the real value of RD. The frequency of failure of a model was determined as the fraction of times (%) when the calculated RD_{MODEL} differed more than four times from the real value of RD.

The robustness of the best model was checked by applying it to data that were randomly selected (in triplicate) from the original data without row replacement (between 2% and 90% of the original dataset). For this purpose, a user-defined function was built under *R* (see Appendix A). The stabilities of the best fit for different randomly selected fractions *f* of the dataset were compared by the success (%) and failure (%) of the fit. Additionally, the robustness of the best model was evaluated by developing it using the first half of the available data as a training set (10/03/1978–20/06/1979). The training set should be large enough to include a wide range of data for the RD. The developed model was then applied to the remainder of the data as a validation set (21/06/1979–30/09/1980).

4 Results and discussion

The average RD, calculated from the available data for the studied period, was around 62 $m^3 \cdot s^{-1}$. This value is similar to the average RD of Vouga River downstream, reported in previous research (Dias et al., 2007). Therefore, the sum of the RD from the two available hydrological stations (the Vouga River and its tributary, Agueda River) accurately represents the downstream RD of the Vouga River. The other small tributaries (Fig. 1) have negligible influence on RD. During the studied period, RD varied from 0.44 $m^3 \cdot s^{-1}$ (14/09/1979) to 1491 $m^3 \cdot s^{-1}$ (09/02/1979). A huge variation of the RD data was observed even on a day to day basis. For a given day *i*, a difference can be defined as $\Delta RD_i = RD_{i+1} - RD_i$, and for the studied dataset, it varied from -400 to $+736$ $m^3 \cdot s^{-1}$.

4.1 Models without temperature

For the studied period, the determination coefficients R^2 of RD with *P*1 (0.323), *P*3 (0.531), *P*5 (0.607) and with *P*2_31 (0.272), although significant, were relatively low. Monthly river runoff has been found to be correlated with monthly precipitation for Yangtze River (Gemmer et al., 2008). Even in that case, the runoff could be influenced by human activities, strongly decreasing its correlation with monthly precipitation. It is even more difficult to find a model for shorter time scale variations, for example to estimate the daily values of RD.

Preliminary examination of the data showed that a single precipitation term, for example accounting for recent precipitation events, *P*5, could predict very high values of RD caused by intense recent precipitation but failed to predict RD for low values of *P*5. Conversely, a precipitation term averaged for longer periods, such as *P*2_31, could account only for very low levels of RD and showed larger variations for higher RD (results not shown). Therefore, in order to reliably estimate the daily RD values using only one precipitation term is not enough. In the present case, the precipitation history was separated into P_R and P_{NR} and both explanatory precipitation variables were used in the multiple regression models.

The coefficients of model 1 are presented in Table 1. It was not possible to simplify the model as all coefficients were statistically significant. The single effect coefficients a_1 (accounting for P_R) and a_2 (accounting for P_{NR}) were all positive and depended on the relative length of time for which P_R (1, 3, or 5 days) and P_{NR} (10 or 30 days) were calculated. The interaction coefficients $a_{1,2}$ were negative and increased (in absolute value) for longer precipitation histories. A significant interaction means that the effect the P_R has on RD will depend on P_{NR} and the model will be difficult to interpret. The R'^2 shows that the choice of longer precipitation history is crucial for the model

Table 1 Significant ($p < 0.05$) regression coefficients (Eq. (1)), Box-Cox parameter λ and adjusted R'^2 for model 1

Fit	$P1$; $P2\_11$[a)]	$P1$; $P2\_31$[b)]	$P3$; $P4\_13$[c)]	$P3$; $P4\_33$[d)]	$P5$; $P6\_15$[e)]	$P5$; $P6\_35$[f)]
a_0	2.470	1.475	2.235	1.344	2.025	1.159
a_1	0.116	0.105	0.254	0.216	0.358	0.298
a_2	0.419	0.363	0.347	0.370	0.303	0.355
$a_{1,2}$	-2.38×10^{-3}	-3.59×10^{-3}	-4.41×10^{-3}	-6.35×10^{-3}	-6.83×10^{-3}	-1.05×10^{-2}
λ	0.2626	0.1414	0.2626	0.1818	0.2626	0.1818
R'^2	0.681	0.673	0.733	0.738	0.752	0.775

[a)] $P_R = P1$, $P_{NR} = P2\_11$; [b)] $P_R = P1$, $P_{NR} = P2\_31$; [c)] $P_R = P3$, $P_{NR} = P4\_13$; [d)] $P_R = P3$, $P_{NR} = P4\_33$; [e)] $P_R = P5$, $P_{NR} = P6\_15$; [f)] $P_R = P5$, $P_{NR} = P6\_35$. $P_R = P_i$ ($i = 1, 3, 5$) are recent daily precipitations P_i, averaged for periods of one, three or five days before the day i; P_{NR} are non-recent daily precipitations averaged for ten or thirty day periods prior to each P_R period.

performances, especially the selection of longer time periods for the calculation of P_R.

All statistically significant coefficients of model 2 are presented in Table 2. The interaction coefficient $a_{1,2}$ was either not significant ($p > 0.1$) for shorter precipitation histories or much smaller compared with the interaction term of model 1 (Table 1). The decreased interaction effect allowed for better physical interpretation of model 2. The simple effect coefficients were positive and determined by the relative length of short and long time periods considered in the model. They account for the effect of P_R and P_{NR} on increasing RD. The coefficients for the quadratic effects $a_{1,1}$ and $a_{2,2}$ were all negative and were also determined by the length of the time periods considered. The absolute values of $a_{1,1}$ and $a_{2,2}$ increased at high values of a_1 and a_2, respectively. Therefore, the RD was determined by both linear effect of P_R and P_{NR} (accounting for increase of RD) and curvature effect of P_R and P_{NR} (accounting for decrease of RD). The R'^2 shows that model 2 was better when larger time periods were used for the calculation of both P_R and P_{NR}.

A comparison between R'^2 of model 1 (0.681–0.775) and model 2 (0.690–0.852) shows a notable improvement in model 2. The frequency of success (%) and failure (%) to predict RD of models 1 and 2 are presented in Figs. 2(a) and 2(b). The most significant improvement in model 2, in comparison with model 1, was observed when longer time periods were used for the calculation of P_{NR}. The highest success frequency was obtained for $P_R = P5$ and $P_{NR} = P6\_35$ (49.0% and 69.3% for models 1 and 2, respectively). The lowest failure frequency was also obtained for $P_R = P5$ and $P_{NR} = P6\_35$ (8.8% and 3.6% for models 1 and 2, respectively).

In order to find the limitations of the models, an analysis of all failure situations was carried out. The ratio between RD_{MODEL} and actual RD for all cases of failure, as a function of RD and ΔRD_i is shown in Fig. 2(c). The datasets showing the best results for both models ($P_R = P5$, $P_{NR} = P6\_35$) were used in this calculation. The failure of model 1, when the calculated values RD_{MODEL} were much lower than the real values (negative errors) was observed for average to high values of RD (16–437 $m^3 \cdot s^{-1}$). However, the main problem of model 1 is that it fails to predict very low values of RD (< 10 $m^3 \cdot s^{-1}$), when the calculated RD_{MODEL} was much higher than the real RD (positive errors). This situation occurred for ΔRD_i close to 0, most often found in the dry season. The addition of quadratic term of precipitation (model 2) significantly improved the prediction of very low RD both by decreasing failure frequency and by reducing the deviation of calculated RD_{MODEL} from the actual values of RD (Fig. 2(c)).

4.2 Models including the temperature

When T was included, but without a curvature effect of the

Table 2 Significant ($p < 0.05$) regression coefficients (Eq. (2)), Box-Cox parameter λ and adjusted R'^2 for model 2

Fit	$P1$; $P2\_11$[a)]	$P1$; $P2\_31$[b)]	$P3$; $P4\_13$[c)]	$P3$; $P4\_33$[d)]	$P5$; $P6\_15$[e)]	$P5$; $P6\_35$[f)]
a_0	2.022	0.5350	1.877	0.4490	1.5830	0.3427
a_1	7.23×10^{-3}	6.37×10^{-2}	0.222	0.140	0.310	0.226
a_2	0.550	0.681	0.566	0.698	0.463	0.722
$a_{1,2}$	–	–	–	-1.35×10^{-3}	-2.44×10^{-3}	-3.67×10^{-3}
$a_{1,1}$	–	-3.54×10^{-4}	-7.92×10^{-4}	-7.13×10^{-4}	-1.81×10^{-3}	-1.05×10^{-3}
$a_{2,2}$	-7.99×10^{-3}	-2.17×10^{-2}	-9.68×10^{-3}	-2.23×10^{-2}	-8.49×10^{-3}	-2.28×10^{-2}
λ	0.2222	0.0606	0.2626	0.1010	0.2222	0.1414
R'^2	0.690	0.789	0.763	0.826	0.768	0.852

[a)] $P_R = P1$, $P_{NR} = P2\_11$; [b)] $P_R = P1$, $P_{NR} = P2\_31$; [c)] $P_R = P3$, $P_{NR} = P4\_13$; [d)] $P_R = P3$, $P_{NR} = P4\_33$; [e)] $P_R = P5$, $P_{NR} = P6\_15$; [f)] $P_R = P5$, $P_{NR} = P6\_35$. $P_R = P_i$ ($i = 1, 3, 5$) are recent daily precipitations P_i, averaged for periods of one, three or five days before the day i; P_{NR} are non-recent daily precipitations averaged for ten or thirty day periods prior to each P_R period.

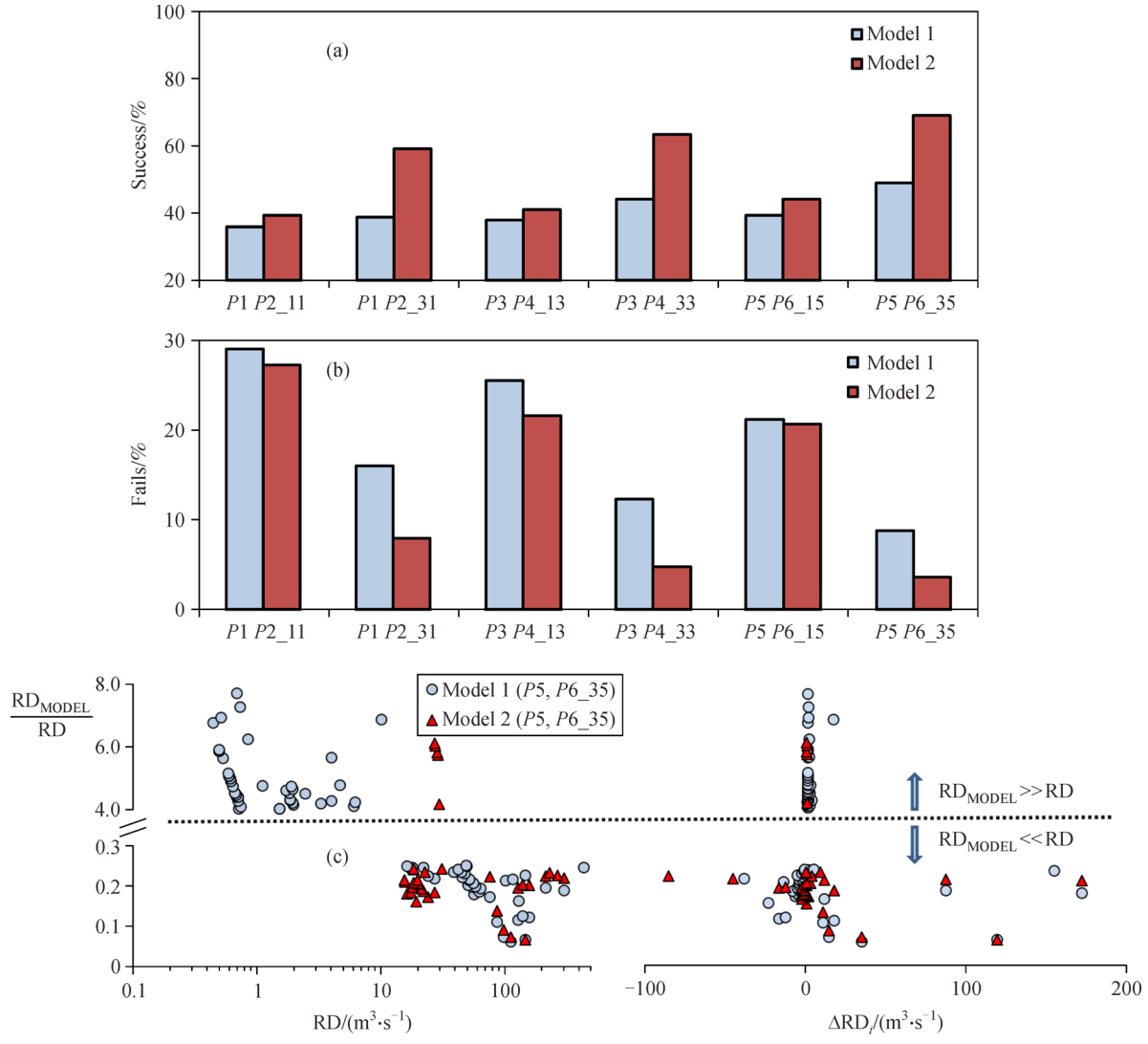

Fig. 2 Success (a) and failure (b) frequencies for RD estimation (models without temperature). Ratio (c) between RD_{MODEL} and actual RD for failure of the best fit (*P*5, *P*6_35) as a function of RD and $\Delta RD_i = RD_{i+1} - RD_i$. $P_R = P_i$ (i = 1, 3, 5) are recent daily precipitations P_i, averaged for periods of one, three or five days before the day i; P_{NR} are non-recent daily precipitations averaged for ten (*P*2_11, *P*4_13, *P*6_15) or thirty day (*P*2_31, *P*4_33, or *P*6_35) periods prior to each P_R period. The frequency of success (%) shows how often the calculated RD_{MODEL} differed from the real value of RD no more than twice. The frequency of failure (%) shows how often the calculated RD_{MODEL} differed from the real value of RD more than four times.

precipitation terms P_R and P_{NR} (model 3, Table 3), the results were better than those obtained by either models 1 or 2 (section 4.1). The improvement of the models when T was included in the calculation was surprising as the temperature data that was readily available were monthly averages. It has been reported that the regression of precipitation and monthly values of evaporation led to more accurate results for RD than estimating from actual evaporation (Hiscock et al., 2001). The R'^2s for model 3 were in the range 0.837 to 0.879 and were not as dependent on the right choice of P_R and P_{NR} datasets as in the previous models. Nevertheless, the R'^2 of model 3 increased when higher time periods were used for the calculation of P_R. The strong interactions found in model 3 constitute its main drawback, as they preclude the possibility of any physical explanation. The interactions tended to be more complex when using a longer precipitation history. The coefficients a_3, accounting for the temperature effects, are always negative, which indicates a lower RD at high temperatures. The coefficients of the main effects of precipitation terms P_R and P_{NR}, while statistically significant, had no physical meaning and were sometimes negative.

When curvature effects of precipitation P_R and P_{NR} were also included (model 4, Table 4) losses from the P_{NR} ($a_{2,2}$) were always significant. The R'^2 was in a narrow range of 0.846–0.888. However, as in model 3, it depended more on the time period chosen to calculate P_R. Again, the strong

Table 3 Significant ($p < 0.05$) regression coefficients (Eq. (3)), Box-Cox parameter λ and adjusted R'^2 for model 3

Fit	$P1$; $P2\_11$; T[a)]	$P1$; $P2\_31$; T[b)]	$P3$; $P4\_13$; T[c)]	$P3$; $P4\_33$; T[d)]	$P5$; $P6\_15$; T[e)]	$P5$; $P6\_35$; T[f)]
a_0	9.395	6.393	8.924	6.782	8.861	6.487
a_1	3.28×10^{-2}	3.45×10^{-2}	0.150	–	0.152	–
a_2	–	–0.284	-4.81×10^{-2}	–0.352	–0.125	–0.358
a_3	–0.426	–0.301	–0.403	–0.327	–0.405	–0.316
$a_{1,2}$	9.46×10^{-4}	3.63×10^{-3}	–	1.61×10^{-2}	6.23×10^{-3}	1.93×10^{-2}
$a_{1,3}$	–	–	-3.93×10^{-3}	6.67×10^{-3}		1.02×10^{-2}
$a_{2,3}$	1.56×10^{-2}	3.74×10^{-2}	1.68×10^{-2}	4.31×10^{-2}	2.06×10^{-2}	4.34×10^{-2}
$a_{1,2,3}$	–	-3.69×10^{-4}	–	-1.50×10^{-3}	-6.09×10^{-4}	-1.96×10^{-3}
λ	0.1818	0.0202	0.1818	0.0606	0.1818	0.0606
R'^2	0.852	0.837	0.870	0.862	0.874	0.879

[a)] $P_R = P1$, $P_{NR} = P2\_11$; [b)] $P_R = P1$, $P_{NR} = P2\_31$; [c)] $P_R = P3$, $P_{NR} = P4\_13$; [d)] $P_R = P3$, $P_{NR} = P4\_33$; [e)] $P_R = P5$, $P_{NR} = P6\_15$; [f)] $P_R = P5$, $P_{NR} = P6\_35$. $P_R = P_i$ ($i = 1, 3, 5$) are recent daily precipitations P_i, averaged for periods of one, three or five days before the day i; P_{NR} are non-recent daily precipitations averaged for ten or thirty day periods prior to each P_R period.

Table 4 Significant ($p < 0.05$) regression coefficients (Eq. (4)), Box-Cox parameter λ and adjusted R'^2 for model 4

Fit	$P1$; $P2\_11$; T[a)]	$P1$; $P2\_31$; T[b)]	$P3$; $P4\_13$; T[c)]	$P3$; $P4\_33$; T[d)]	$P5$; $P6\_15$; T[e)]	$P5$; $P6\_35$; T[f)]
a_0	8.820	5.179	8.488	5.314	8.256	5.009
a_1	2.70×10^{-2}	4.06×10^{-2}	7.91×10^{-2}	–	8.29×10^{-2}	–
a_2	0.158	–	0.134	–	7.94×10^{-2}	–
a_3	–0.395	–0.241	–0.382	–0.254	–0.376	–0.243
$a_{1,2}$	1.40×10^{-3}	–	1.89×10^{-3}	1.61×10^{-2}	8.70×10^{-3}	2.07×10^{-2}
$a_{1,3}$	–	–	–	6.49×10^{-3}	4.89×10^{-3}	1.02×10^{-2}
$a_{2,3}$	7.72×10^{-3}	2.44×10^{-2}	8.72×10^{-3}	2.76×10^{-2}	1.17×10^{-2}	2.76×10^{-2}
$a_{1,2,3}$	–	–	–	-1.48×10^{-3}	-7.56×10^{-4}	-2.09×10^{-3}
$a_{1,1}$	–	-1.76×10^{-4}	–	–	–	–
$a_{2,2}$	-2.56×10^{-3}	-6.65×10^{-3}	-3.87×10^{-3}	-8.53×10^{-3}	-3.87×10^{-3}	-8.69×10^{-3}
λ	0.1818	0.0202	0.1818	0.0606	0.1818	0.0606
R'^2	0.854	0.846	0.876	0.872	0.881	0.888

[a)] $P_R = P1$, $P_{NR} = P2\_11$; [b)] $P_R = P1$, $P_{NR} = P2\_31$; [c)] $P_R = P3$, $P_{NR} = P4\_13$; [d)] $P_R = P3$, $P_{NR} = P4\_33$; [e)] $P_R = P5$, $P_{NR} = P6\_15$; [f)] $P_R = P5$, $P_{NR} = P6\_35$. $P_R = P_i$ ($i = 1, 3, 5$) are recent daily precipitations P_i, averaged for periods of one, three or five days before the day i; P_{NR} are non-recent daily precipitations averaged for ten or thirty day periods prior to each P_R period.

interactions did not allow physical interpretation of the model.

The success and failure frequencies (%) of models 3 and 4 to predict RD are presented in Figs. 3(a) and 3(b). Similar to the models without temperature, the most significant improvement of model 4, compared to model 3, was when longer time periods were used for the calculation of P_{NR}. The highest success frequency was obtained for $P_R = P5$ and $P_{NR} = P6\_35$ (79.1% for both models 3 and 4). The lowest failure frequency was also obtained for $P_R = P5$ and $P_{NR} = P6\_35$ (4.3% and 2.4% for models 3 and 4, respectively). However, only a small improvement was observed in model 4 (Fig. 3) when curvature effects of precipitation terms were introduced. As discussed in section 4.1, the introduction of curvature effects of precipitation for models without temperature led to a very significant improvement (Fig. 2). Therefore, the quadratic effect of P_R and P_{NR}, accounting for the precipitation lost by evaporation or retention by the soils, discussed for model 2 (section 4.1) was due to effects of T on RD. For the Blue River in Oklahoma, with increased soil dryness and water evaporation during the hot season, a significant reduction in the mean monthly RD for the same monthly precipitation input was identified (Gourley and Vieux, 2006).

For all cases of failure, the ratio between RD_{MODEL} and actual RD (calculated for $P_R = P5$, $P_{NR} = P6\_35$) is shown in Fig. 3(c) as a function of RD and ΔRD_i. The negative errors of model 3 were observed for low to average values of RD (4.7–158 $m^3 \cdot s^{-1}$). Sometimes, in the dry season, model 3 fails to predict ($RD_{MODEL} >> RD$, positive errors) relatively low values of RD (1.3–10.8 $m^3 \cdot s^{-1}$) with ΔRD_i close to 0. Again, the addition of quadratic terms of precipitation (model 4) significantly reduced primarily the

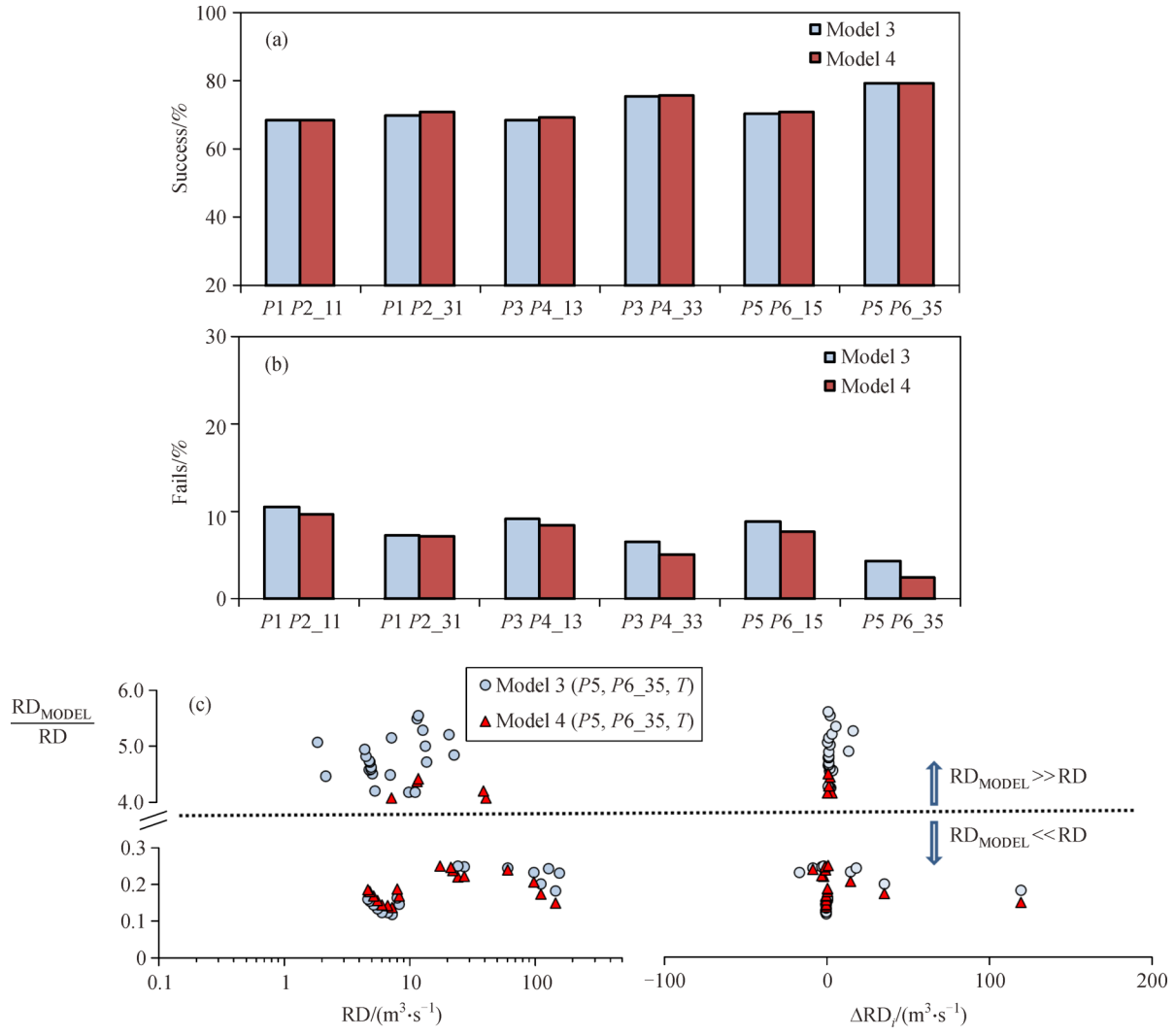

Fig. 3 Success (a) and failure (b) frequencies for RD estimation (models including temperature). Ratio (c) between RD_{MODEL} and actual RD for failure of the best fit (*P*5, *P*6_35, *T*) as a function of RD and $\Delta RD_i = RD_{i+1} - RD_i$. $P_R = P_i$ ($i = 1, 3, 5$) are recent daily precipitations P_i, averaged for periods of one, three or five days before the day i; P_{NR} are non-recent daily precipitations averaged for ten (*P*2_11, *P*4_13, *P*6_15) or thirty day (*P*2_31, *P*4_33, or *P*6_35) periods prior to each P_R period. The frequency of success (%) shows how often the calculated RD_{MODEL} differed from the real value of RD no more than twice. The frequency of failure (%) shows how often the calculated RD_{MODEL} differed from the real value of RD more than four times.

positive errors both by decreasing the failure frequency and by reducing the deviation of calculated RD_{MODEL} from the actual values of RD (Fig. 3(c)).

4.3 Robustness

The success and failure frequencies (%) for different random selections of fraction *f* from the original datasets, for model 4 (*T*, P_R= *P*5, P_{NR}= *P*6_35) are presented in Figs. 4(a) and 4(b), respectively. Model 4 worked reasonably well after removing up to 60% of the data. As larger fractions of *f* were removed from the data, the prediction by model 4 gradually worsened, as demonstrated by the decrease of its success and increase of its failure frequencies (Figs. 4(a) and 4(b)). Additionally, the error bars became larger, showing the increasing instability of the prediction which, additionally, depends upon the random selection of the data. A comparison of RD calculated with all data ($f = 1$) with RD calculated using three replicates of randomly selected data ($f = 0.06$) is presented in Fig. 4(c). The largest instability was found when the RD, as a function of time, had maximum or minimum values. The same was observed for other fractions $f < 0.4$ of randomly selected data (results not shown).

A comparison of the actual variation in RD over time with that predicted using the best model in this study, model 4 (*T*, P_R= *P*5, P_{NR}= *P*6_35) is presented in Fig. 5. In general, the variation in RD over time is predicted reasonably well by the model. However, even the best

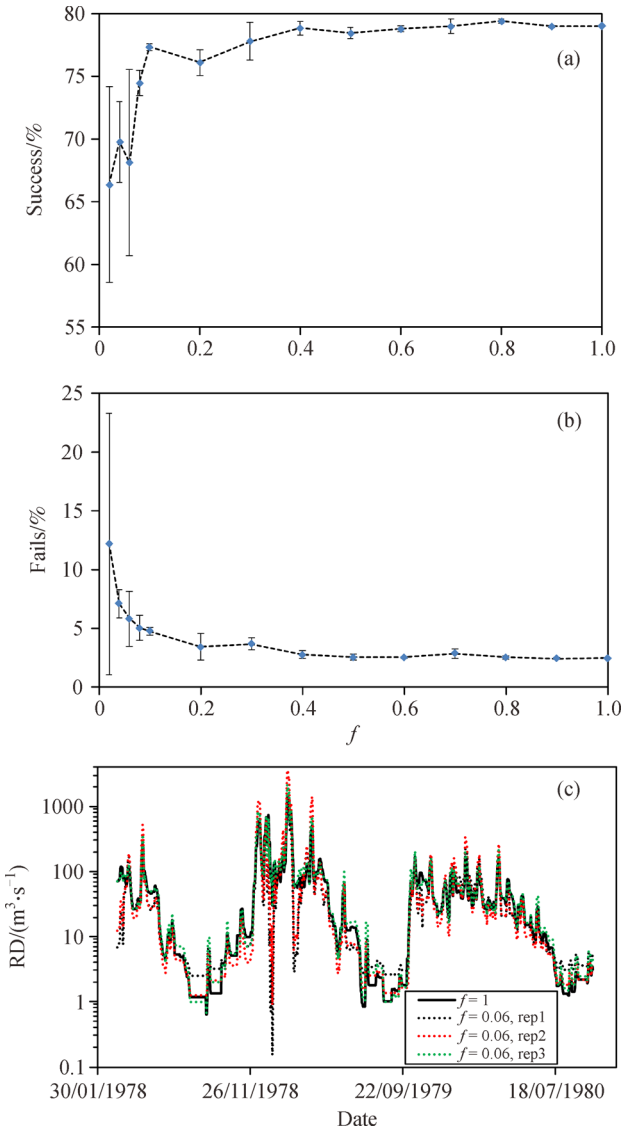

Fig. 4 Success (a) and failure (b) frequencies for RD estimated by model 4 (P_R= *P*5, P_{NR}= *P*6_35, *T*) for randomly selected fractions *f* of the original dataset; comparison (c) of the time dependence of RD estimated by model 4 (P_R= *P*5, P_{NR}= *P*6_35, *T*) using all data (*f* = 1) and three replicates of randomly selected data (*f* = 0.06). The frequency of success (%) shows how often the calculated RD_{MODEL} differed from the real value of RD no more than twice. The frequency of failure (%) shows how often the calculated RD_{MODEL} differed from the real value of RD more than four times. P_R= *P*5 are recent daily precipitations P_i, averaged for periods of five days before the day *i*; P_{NR}= *P*6_35 are non-recent daily precipitations averaged for thirty day periods prior to the P_R period.

model in this study calculated slightly higher values than the real values for low RD. Additionally, model 4 was developed for the first half of the study period (training set). Subsequently, the developed model was applied to the validation set and closely followed the real data (Fig. 5) with success and failure frequencies of 72.9% and 5.7%, respectively.

5 Conclusions

Very simple and robust models to calculate average daily discharge using only readily available data of precipitation and average monthly air temperature were developed and compared, using the case study of the Vouga River that discharges in the Aveiro Lagoon. Since the Vouga River is not influenced by melting of ice and snow, the proposed models should be applied only for rivers not dependent on these processes. These models were based on the analysis of historic datasets of both RD and meteorological data. Interestingly, input data from only one meteorological station was enough to estimate the RD. A separation of the precipitation history into recent and non-recent events (before the day for which RD is to be calculated) was found to be essential. For models which only take into account the precipitation terms, it was also crucial to

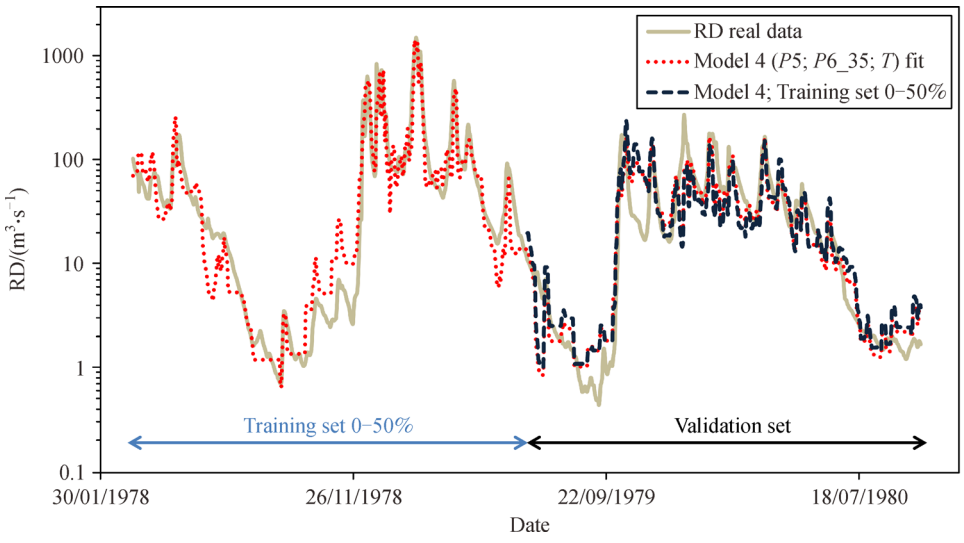

Fig. 5 Variation over time of the real RD (continuous line) and of the RD estimated by model 4 fit (red dots, all data, P_R= *P*5, P_{NR}= *P*6_35, *T*). The same model was also developed using the training set 10/03/1978–20/06/1979 and applied for the remaining data as the validation set (black hyphens). P_R= *P*5 are recent daily precipitations P_i, averaged for periods of five days before the day *i*; P_{NR}= *P*6_35 are non-recent daily precipitations averaged for thirty day periods prior to the P_R period.

include curvature effects representing water loss, for example by evaporation and retention by the soils. Better models were generally found when *T* was also included, but the physical meaning was lost due to the importance of the interaction terms. The developed models with historical data can be a useful tool for estimating RD at the present moment using available meteorological data at the same site (for example from http://www.wunderground.com). However, this estimation should be considered with caution in case there are major changes of the land use in the drainage area. The estimated data of RD could be used for subsequent modeling of biogeochemical processes in estuaries and rivers.

Acknowledgements This research was partially supported by the Strategic Funding UID/Multi/04423/2013 through national funds provided by FCT – Foundation for Science and Technology and European Regional Development Fund (ERDF), in the framework of the programme PT2020. T. Stoichev is grateful to FCT for his fellowship (SFRH/BPD/88675/2012), co-financed by Programa Operacional Potencial Humano (POPH) / Fundo Social Europeu (FSE). J. Espinha Marques and R. Moura acknowledge the funding provided by the Institute of Earth Sciences (ICT), under contract with FCT.

References

Achleitner S, Schöber J, Rinderer M, Leonhardt G, Schöberl F, Kirnbauer R, Schönlaub H (2012). Analyzing the operational performance of the hydrological models in an alpine flood forecasting system. J Hydrol (Amst), 412–413: 90–100

AEMET-IM (2011). Iberian Climate Atlas- Air temperature and precipitation (1971–2000). Ministerio de Medio Ambiente y Medio Rural y Marino (Spain), Instituto de Meteorologia (Portugal)

Agroconsultores and Geometral (1995). Carta dos solos e da aptidão da terra do Entre-Douro e Minho [Map of Soils and Land Suitability of Entre-Douro and Minho]. Lisbon: DRAEDM

Bellanger L, Tomassone R (2014). Exploration de données et méthodes statistiques: data analysis & data mining avec le logiciel R [Data Exploration and Statistical Methods: Data Analysis & Data Mining Using R]. Paris: Ellipses

Beven K J (1989). Changing ideas in hydrology: the case of physically-based models. J Hydrol (Amst), 105(1–2): 157–172

Beven K J (2012). Rainfall–Runoff Modeling: The Primer. Chichester: Wiley

Brath A, Montanari A, Toth E (2004). Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model. J Hydrol (Amst), 291(3–4): 232–253

Cerejo M, Dias J M (2007). Tidal transport and dispersal of marine toxic microalgae in a shallow, temperate coastal lagoon. Mar Environ Res, 63(4): 313–340

Crawley M J (2007). The R book. Chichester: Wiley

Dias J M, Abrantes I, Rocha F (2007). Suspended particulate matter sources and residence time in a mesotidal lagoon. J Coast Res, 50 (Special issue): 1034–1039

Dias J M, Lopes J F, Dekeyser I (1999). Hydrological characterisation of Ria de Aveiro, Portugal, in early summer. Oceanol Acta, 22(5): 473–485

Du J, Xie S, Xu Y, Xu C, Singh V P (2007). Development and testing of a simple physically-based distributed rainfall-runoff model for storm runoff simulation in humid forested basins. J Hydrol (Amst), 336(3–4): 334–346

ESBN (2005). Soil Atlas of Europe. Luxembourg: European Commission

Espinha Marques J, Samper J, Pisani B, Alvares D, Carvalho J M, Chaminé H I, Marques J M, Vieira G T, Mora C, Sodré Borges F (2011). Evaluation of water resources in a high-mountain basin in

Serra da Estrela, Central Portugal, using a semi-distributed hydrological model. Environmental Earth Sciences, 62(6): 1219–1234

Gemmer M, Jiang T, Su B, Kundzewicz Z W (2008). Seasonal precipitation changes in the wet season and their influence on flood/drought hazards in the Yangtze River Basin, China. Quat Int, 186(1): 12–21

Gourley J J, Vieux B E (2006). A method for identifying sources of model uncertainty in rainfall-runoff simulations. J Hydrol (Amst), 327(1–2): 68–80

Hiscock K M, Lister D H, Boar R R, Green F M L (2001). An integrated assessment of long-term changes in the hydrology of three lowland rivers in eastern England. J Environ Manage, 61(3): 195–214

Hurkmans R T W L, de Moel H, Aerts J C J H, Troch P A (2008). Water balance versus land surface model in the simulation of Rhine river discharges. Water Resour Res, 44(1): W01418

IPMA (2015). Normais climatológicas 1971–2000 [Weather and Climate 1971–2000]. Retrieved from: http://www.ipma.pt/en/oclima/normais.clima/

Muthukrishnan S, Harbor J, Lim K J, Engel B A (2006). Calibration of a simple rainfall-runoff model for long-term hydrological impact evaluation. URISA Journal, 18(2): 35–42

Oliveira J T, Pereira E, Ramalho M, Antunes M T, Monteiro J H (1992). Carta Geológica de Portugal 1/500 000 [Geological Map of Portugal 1/500 000]. 5th ed. Lisbon: Serviços Geológicos de Portugal

R Core Team (2014). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/

Rogado N J O, Batalha J F C S, Simões J J M F, Ribeiro L M (1992). *Esboço duma carta de solos da Região de Aveiro na escala 1/100 000* [Project of a soil map of Aveiro region on the scale 1/100 000]. Coimbra (Portugal): DRABL

Sen Z (2010). Fuzzy Logic and Hydrological Modelling. Boca Raton: Taylor and Francis

USGS (2009). Shuttle Radar Topography Mission 3-arc second data (version 2.1). Retrieved from: http://dds.cr.usgs.gov/srtm/version2_1/SRTM3/

Van der Weijden C H, Pacheco F A L (2006). Hydrogeochemistry in the Vouga River basin (central Portugal): pollution and chemical weathering. Appl Geochem, 21(4): 580–613

Vilaysane B, Takara K, Luo P, Akkharath I, Duan W (2015). Hydrological stream flow modelling for calibration and uncertainty analysis using SWAT model in the Xedone river basin, Lao PDR. Procedia Environ Sci, 28: 380–390

Xu H, Taylor R G, Xu Y (2011). Quantifying uncertainty in the impacts of climate change on river discharge in sub-catchments of the Yangtze and Yellow River Basins, China. Hydrol Earth Syst Sci, 15 (1): 333–344

Appendix A

A user-built function, called “select1” was developed under R for random selection without replacement of fraction f < 1 (from “fbegin” to “fend” increasing with step “fstep”) of the original data from the dataframe x:

```
select1 < -function(x, fbegin, fend, fstep){
rounded < -function(n) floor(n + 0.5)
f < -fbegin
sel < -x[sample(1:length(x[,1]),rounded(f*length(x[,1]))),]
repeat{
t < -max(f)
if (t > = fend) break
sel1 < -x[sample(1:length(x[,1]),rounded((t + fstep)*length(x[,1]))),]
f < -c(f, (t + fstep))
sel < -list(sel, sel1)}
return(list(f, sel))}
```